CNRC-TMT:
Second Language Writing Assistant System Description

Cyril Goutte **Michel Simard** **Marine Carpuat**
National Research Council Canada
Multilingual Text Processing
1200 Montreal Road, Ottawa, Ontario K1A 0R6, Canada
FirstName.LastName@nrc.ca

Abstract

We describe the system entered by the National Research Council Canada in the SemEval-2014 L2 writing assistant task. Our system relies on a standard Phrase-Based Statistical Machine Translation trained on generic, publicly available data. Translations are produced by taking the already translated part of the sentence as fixed context. We show that translation systems can address the L2 writing assistant task, reaching out-of-five word-based accuracy above 80 percent for 3 out of 4 language pairs. We also present a brief analysis of remaining errors.

1 Introduction

The Semeval L2 writing assistant task simulates the situation of an L2 language learner trying to translate a L1 fragment in a L2 context. This is clearly motivated by a L2 language learning scenario.

However, a very similar scenario can be encountered in Computer-Aided Translation. Translation memories retrieve from a large corpus of already translated documents the source segments that best match a new sentence to be translated. If an exact source match is found, the corresponding target translation can be expected to be suitable with little or no post-editing. However, when only approximate matches are found, post-editing will typically be required to adapt the target side of the partially matching source segment to the source sentence under consideration. It is possible to automate this process: standard string matching algorithms and word alignment techniques can be used to locate the parts of the source segment that do not match the sentence to translate, and from there the parts of the target segment that need to be modified (Biçici and Dymetman, 2008; Simard and Isabelle, 2009; Koehn and Senellart, 2010). The task of translating a L1 fragment in L2 context therefore has much broader application than language learning. This motivation also provides a clear link of this task to the Machine Translation setting. There are also connections to the code-switching and mixed language translation problems (Fung et al., 1999).

In our work, we therefore investigate the use of a standard Phrase-Based Statistical Machine Translation (SMT) system to translate L1 fragments in L2 context. In the next section, we describe the SMT system that we used in our submission. We then describe the corpora used to train the SMT engine (Section 3), and our results on the trial and test data, as well as a short error analysis (Section 4).

section

2 System Description

The core Machine Translation engine used for all our submissions is Portage (Larkin et al., 2010), the NRC's phrase-based SMT system. Given a suitably trained SMT system, the Task 5 input is processed as follows. For each sentence with an L1 fragment to translate, the already translated parts are set as left and right context. The L1 fragment in L2 context is sent to the decoder. The output is a full sentence translation that ensures 1) that the context is left untouched, and 2) that the L1 fragment is translated in a way that fits with the L2 context.

We now describe the key components of the MT system (language, translation and reordering models), as well as the decoding and parameter tuning.

Translation Models We use a single static phrase table including phrase pairs extracted from the symmetrized HMM word-alignment learned

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 192–197,
Dublin, Ireland, August 23-24, 2014.

on the entire training data. The phrase table contains four features per phrase pair: lexical estimates of the forward and backward probabilities obtained either by relative frequencies or using the method of Zens and Ney (2004). These estimates are derived by summing counts over all possible alignments. This yields four corresponding parameters in the log-linear model.

Reordering Models We use standard reordering models: a distance-based distortion feature, as well as a lexicalized distortion model (Tillmann, 2004; Koehn et al., 2005). For each phrase pair, the orientation counts required for the lexicalized distortion model are computed using HMM word-alignment on the full training corpora. We estimate lexicalized probabilities for monotone, swap, and discontinuous ordering with respect to the previous and following target phrase. This results in a total of 6 feature values per phrase pair, in addition to the distance-based distortion feature, hence seven parameters to tune in the log-linear model.

Language Models When translating L1 fragments in L2 context, the L2 language model (LM) is particularly important as it is the only component of the SMT system that scores how well the translation of the L1 fragment fits in the existing L2 context. We test two different LM configurations. The first of these (*run1*) uses a single static LM: a standard 4-gram, estimated using Kneser-Ney smoothing (Kneser and Ney, 1995) on the target side of the bilingual corpora used for training the translation models. In the second configuration (*run2*), in order to further adapt the translations to the test domain, a smaller LM trained on the L2 contexts of the test data is combined to the training corpus LM in a linear mixture model (Foster and Kuhn, 2007). The linear mixture weights are estimated on the L2 context of each test set in a cross-validation fashion.

Decoding Algorithm and Parameter Tuning Decoding uses the cube-pruning algorithm (Huang and Chiang, 2007) with a 7-word distortion limit. Log-linear parameter tuning is performed using a lattice-based batch version of MIRA (Cherry and Foster, 2012).

3 Data

SMT systems require large amounts of data to estimate model parameters. In addition, translation performance largely depends on having in-domain data to train on. As we had no information on the domain of the test data for Task 5, we chose to rely on general purpose publicly available data. Our main corpus is Europarl (Koehn, 2005), which is available for all 4 language pairs of the evaluation. As Europarl covers parliamentary proceedings, we added some news and commentary (henceforth "News") data provided for the 2013 workshop on Machine Translation shared task (Bojar et al., 2013) for language pairs other than nl-en. In all cases, we extracted from the corpus a tuning ("dev") set of around 2000 sentence pairs. Statistics for the training data are given in Table 1.

		Europarl	News	Total
en-de	train	1904k	177k	2081k
	dev	-	2000	2000
en-es	train	1959k	174k	2133k
	dev	-	2000	2000
fr-en	train	2002k	157k	2158k
	dev	-	2000	2000
nl-en	train	1974k	-	1974k
	dev	1984	-	1984

Table 1: Number of training segments for each language pair.

The trial and test data each consist of 500 sentences with L1 fragments in L2 context provided by the organizers. As the trial data came from Europarl, we filtered our training corpora in order to remove close matches and avoid training on the trial data (Table 1 takes this into account).

All translation systems were trained on lowercased data, and predictions were recased using a standard (LM-based) truecasing approach.

4 Experimental Results

4.1 Results on Trial and Simulated Data

Our first evaluation was performed on the trial data provided by the Task 5 organizers. Each example was translated in context by two systems:

run1: Baseline, non-adapted system (marked **1** below);

run2: Linear LM mixture adaptation, using a context LM (marked **2** below).

Table 2 shows that our *run1* system already yields high performance on the trial data, while

	W@1	F@1	W@5	F@5	+BLEU
en-de1	78.1	77.0	95.6	94.8	12.4
en-de2	79.8	79.0	95.8	95.0	12.6
en-es1	81.8	80.2	97.7	97.2	12.1
en-es2	84.3	83.2	97.7	97.2	12.5
fr-en1	84.4	83.6	97.1	96.4	11.8
fr-en2	85.9	85.0	97.4	96.6	12.0
nl-en1	83.3	82.0	97.0	96.4	11.8
nl-en2	86.7	86.2	97.5	97.0	12.1

Table 2: Trial data performance, from official evaluation script: (W)ord and (F)ragment accuracy at (1) and (5)-best and BLEU score gain.

adapting the language model on the L2 contexts in *run2* provides a clear gain in the top-1 results. That improvement all but disappears when taking into account the best out of five translations (except maybe for nl-en). The BLEU scores[1] are very high (97-98) and the word error rates (not reported) are around 1%, suggesting that the system output almost matches the references. This is no doubt due to the proximity between the trial data and the MT training corpus. Both are fully or mainly drawn from Europarl material.

In order to get a less optimistic estimate of performance, we automatically constructed a number of test examples from the WMT News Commentary development test sets. The L1 source segments and their L2 reference translations were word aligned in both directions using the GIZA++ implementation of IBM4 (Och and Ney, 2003) and the grow-diag-final-and combination heuristic (Koehn et al., 2005). Test instances were created by substituting some L2 fragments with their word-aligned L1 source within L2 reference segments. Since the goal was to select examples that were more ambiguous and harder to translate than the trial data, a subset of interesting L1 phrases was randomly selected among phrases that occured at least 4 times in the training corpus and have a high entropy in the baseline phrase-table. We selected roughly 1000 L1 phrases per language pair. For each occurrence p_1 of these L1 phrases in the news development sets, we identify the shortest L2 phrase p_2 that is consistently aligned with p_1.[2] A new mixed language test example is constructed by replacing p_2 with p_1 in L2 context.

[1] +BLEU in Tables 2-4 is the difference between our system's output and the sentence with untranslated L1 fragment.

	W@1	F@1	W@5	F@5	+BLEU
en-de1	48.0	46.4	70.8	68.7	4.26
en-de2	52.3	50.6	71.0	68.9	4.63
en-es1	47.6	45.2	68.0	65.8	4.12
en-es2	50.0	47.9	67.8	65.5	4.34
fr-en1	50.1	49.2	73.6	71.8	5.18
fr-en2	51.1	49.5	73.1	71.2	5.19

Table 3: News data performance (cf Tab. 2).

Results on that simulated data are given in Table 3. Performance is markedly lower than on the trial data. This is due in part to the fact that the News data is not as close to the training material as the official trial set, and in part to the fact that this automatically extracted data contains imperfect alignments with an unknown (but sizeable) amount of "noise". However, it still appears *run2* consistently provides several points of increase in performance for the top-1 results, over the baseline *run1*. Performance on the 5-best is either unaffected or lower, and the gain in BLEU is much lower than in Table 2 although the resulting BLEU is around 96%.

4.2 Test Results

Official test results provided by the organizers are presented in Table 4. While these results are clearly above what we obtained on the synthetic news data, they fall well below the performance observed on the trial data. This is not unexpected as the trial data is unrealistically close to the training material, while the automatically extracted news data is noisy. What we did not expect, however, is the mediocre performance of LM adaptation (*run2*): while consistently better than *run1* on both trial and news, it is consistently worse on the official test data. This may be due to the fact that test sentences were drawn from different sources[3] such that it does not constitute a homogeneous domain on which we can easily adapt a language model.

For German and Spanish, and to a lesser extent

[2] As usual in phrase-based MT, two phrases are said to be consistently aligned, if there is at least one link between their words and no external links.

[3] According to the task description, the test set is based on "language learning exercises with gaps and cloze-tests, as well as learner corpora with annotated errors".

	W@1	F@1	W@5	F@5	+BLEU
en-de1	71.7	65.7	86.8	83.4	16.6
en-de2	70.2	64.5	86.5	82.8	16.4
en-es1	74.5	66.7	88.7	84.3	17.0
en-es2	73.5	65.1	88.4	83.7	17.5
fr-en1	69.4	55.6	83.9	73.9	10.2
fr-en2	68.6	53.3	83.4	73.1	9.9
nl-en1	61.0	45.0	72.3	60.6	5.03
nl-en2	60.9	44.4	72.1	60.2	5.02

Table 4: Test data performance, from official evaluation results (cf. Table 2).

	OOV's	failed align
en-de	0.002	0.058
en-es	0.010	0.068
fr-en	0.026	0.139
nl-en	0.123	0.261

Table 5: Test data error analysis: *OOV's* is the proportion of all test fragments containing out-of-vocabulary tokens; *failed align* is the proportion of fragments which our system cannot align to any of the reference translations by forced decoding.

for French and Dutch, the BLEU and Word Error Rate (WER) gains are much higher on the test than on the trial data, although the resulting BLEU are around 86-92%. This results from the fact that the amount of L1 material to translate relative to the L2 context was significantly higher on the test data than it was on the trial data (e.g. 17% of words on en-es test versus 7% on trial).

4.3 Error Analysis

On the French and, especially, on the Dutch data, our systems suffer from a high rate of out-of-vocabulary (OOV) source words in the L1 fragments, i.e. words that simply did not appear in our training data (see Table 5). In the case of Dutch, OOV's impose a hard ceiling of 88% on fragment-level accuracy. These problems could possibly be alleviated by using more training data, and incorporating language-specific mechanisms to handle morphology and compounding into the systems.

We also evaluate the proportion of reference target fragments that can not be reached by *forced decoding* (Table 5). Note that to produce trial and test translations, we use standard decoding to predict a translation that maximizes model score given the input. Once we have the reference translation, we use *forced decoding* to try to produce the exact reference given the source fragment and our translation model. In some situations, the correct translations are simply not reachable by our systems, either because some target word has not been observed in training, some part of the correspondence between source and target fragments has not been observed, or the system's word alignment mechanism is unable to account for this correspondence, in whole or in part. Table 5 shows that this happens between 6% ans 26% of cases, which gives a better upper bound on the fragment-level accuracy that our system may achieve. Again, many of these problems could be solved by using more training data.

Freq	Type
77	Incorrect L2 sense chosen
75	Incorrect or mangled syntax
26	Incomplete reference
20	Non-idiomatic translation
13	Out-of-vocab. word in fragment
6	Problematic source fragment
3	Casing error
220	Total

Table 6: Analysis of the types of error on 220 French-English test sentences.

To better understand the behavior of our systems, we manually reviewed 220 sentences where our baseline French-English system did not exactly match any of the references. We annotated several types of errors (Table 6). The most frequent source of errors is incorrect sense (35%), i.e. the system produced a translation of the fragment that may be correct in some setting, but is not the correct sense in that context. Those are presumably the errors of interest in a sense disambiguation setting. A close second (34%) were errors involving incorrect syntax in the fragment translation, which points to limitations of the Statistical MT approach, or to a limited language model.

The last third combines several sources of errors. Most notable in this category are *non-idiomatic translations*, where the system's output was both syntactically correct and understandable, but clearly not fluent (e.g. "take a siesta" for "have a nap"); We also identified a number of cases

where we felt that either the source segment was incorrect (eg "je vais évanouir" instead of "je vais m'évanouir"), or the references were incomplete. Table 7 gives a few examples.

5 Conclusion

We described the systems used for the submissions of the National Research Council Canada to the L2 writing assistant task. We framed the problem as a machine translation task, and used standard statistical machine translation systems trained on publicly available corpora for translating L1 fragments in their L2 context. This approach leverages the strengths of phrase-based statistical machine translation, and therefore performs particularly well when the test examples are close to the training domain. Conversely, it suffers from the inherent weaknesses of phrase-based models, including their inability to generalize beyond seen vocabulary, as well as sense and syntax errors. Overall, we showed that machine translation systems can be used to address the L2 writing assistant task with a high level of accuracy, reaching out-of-five word-based accuracy above 80 percent for 3 out of 4 language pairs.

References

Ergun Biçici and Marc Dymetman. 2008. Dynamic Translation Memory: Using Statistical Machine Translation to Improve Translation Memory Fuzzy Matches. In *Computational Linguistics and Intelligent Text Processing*, pages 454–465. Springer.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria, August.

Colin Cherry and George Foster. 2012. Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436, Montréal, Canada, June.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 128–135, Prague, Czech Republic, June.

Pascale Fung, Xiaohu Liu, and Chi Shun Cheung. 1999. Mixed Language Query Disambiguation. In *Proceedings of ACL'99*, pages 333–340, Maryland, June.

Liang Huang and David Chiang. 2007. Forest Rescoring: Faster Decoding with Integrated Language Models. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 144–151, Prague, Czech Republic, June.

Reinhard Kneser and Hermann Ney. 1995. Improved Backing-off for M-gram Language Modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 181–184. IEEE.

Philipp Koehn and Jean Senellart. 2010. Convergence of Translation Memory and Statistical Machine Translation. In *Proceedings of AMTA Workshop on MT Research and the Translation Industry*, pages 21–31.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. In *Proceedings of IWSLT-2005*, pages 68–75, Pittsburgh, PA.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Machine Translation Summit X*, pages 79–86, Phuket, Thailand, September.

Samuel Larkin, Boxing Chen, George Foster, Ullrich Germann, Éric Joanis, J. Howard Johnson, and Roland Kuhn. 2010. Lessons from NRC's Portage System at WMT 2010. In *5th Workshop on Statistical Machine Translation*, pages 127–132.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–52.

Michel Simard and Pierre Isabelle. 2009. Phrase-based Machine Translation in a Computer-assisted Translation Environment. *Proceedings of the Twelfth Machine Translation Summit (MT Summit XII)*, pages 120–127.

Christoph Tillmann. 2004. A Unigram Orientation Model for Statistical Machine Translation. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Short Papers*, pages 101–104, Boston, Massachusetts, USA, May 2 - May 7.

Richard Zens and Hermann Ney. 2004. Improvements in Phrase-Based Statistical Machine Translation. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 257–264, Boston, Massachusetts, USA, May 2 - May 7.

Incorrect L2 sense:	
In:	My dog usually barks **au facteur** - but look at that , for once , he is being friendly ...
Out:	My dog usually barks **to the factor** - but look at that , for once , he is being friendly ...
Ref:	My dog usually barks **at the postman** - but look at that , for once , he is being friendly ...
In:	Grapes **ne poussent pas** in northern climates , unless one keeps them in a hot-house .
Out:	Grapes **do not push** in northern climates , unless one keeps them in a hot-house .
Ref:	Grapes **do not grow** in northern climates , unless one keeps them in a hot-house .
Missing reference?	
In:	Twenty-two other people **ont été blessées** in the explosion .
Out:	Twenty-two other people **were injured** in the explosion .
Ref:	Twenty-two other people **have been wounded** in the explosion .
Non-idiomatic translation:	
In:	After patiently stalking its prey , the lion makes a **rapide comme l' éclair** charge for the kill .
Out:	After patiently stalking its prey , the lion makes a **rapid as flash** charge for the kill .
Ref:	After patiently stalking its prey , the lion makes a **lightning-fast** charge for the kill .
Problem with input:	
In:	every time I do n't eat for a while and my blood sugar gets low I feel like **je vais évanouir** .
Out:	every time I do n't eat for a while and my blood sugar gets low I feel like **I will evaporate** .
Ref:	every time I do n't eat for a while and my blood sugar gets low I feel like **I 'm going to faint** .

Table 7: Examples errors on French-English.