The CNGL-DCU-Prompsit Translation Systems for WMT13

Raphael Rubino[†], **Antonio Toral**[†], **Santiago Cortés Vaíllo**[⋆],
Jun Xie[§], **Xiaofeng Wu**[‡], **Stephen Doherty**[♭], **Qun Liu**[‡]
[†]NCLT, Dublin City University, Ireland
[⋆]Prompsit Language Engineering, Spain
[§]ICT, Chinese Academy of Sciences, China
[‡,♭]CNGL, Dublin City University, Ireland
[†,‡]{rrubino, atoral, xfwu, qliu}@computing.dcu.ie
[⋆]santiago@prompsit.com
[§]junxie@ict.ac.cn
[♭]stephen.doherty@dcu.ie

Abstract

This paper presents the experiments conducted by the Machine Translation group at DCU and Prompsit Language Engineering for the WMT13 translation task. Three language pairs are considered: Spanish-English and French-English in both directions and German-English in that direction. For the Spanish-English pair, the use of linguistic information to select parallel data is investigated. For the French-English pair, the usefulness of the small in-domain parallel corpus is evaluated, compared to an out-of-domain parallel data sub-sampling method. Finally, for the German-English system, we describe our work in addressing the long distance reordering problem and a system combination strategy.

1 Introduction

This paper presents the experiments conducted by the Machine Translation group at DCU[1] and Prompsit Language Engineering[2] for the WMT13 translation task on three language pairs: Spanish-English, French-English and German-English. For these language pairs, the language and translation models are built using different approaches and datasets, thus presented in this paper in separate sections.

In Section 2, the systems built for the Spanish-English pair in both directions are described. We investigate the use of linguistic information to select parallel data. In Section 3, we present the systems built for the French-English pair in both directions. The usefulness of the small in-domain parallel corpus is evaluated, compared to an out-of-domain parallel data sub-sampling method. In Section 4, for the German-English system, aiming at exploring the long distance reordering problem, we first describe our efforts in a dependency tree-to-string approach, before combining different hierarchical systems with a phrase-based system and show a significant improvement over three baseline systems.

2 Spanish-English

This section describes the experimental setup for the Spanish-English language pair.

2.1 Setting

Our setup uses the MOSES toolkit, version 1.0 (Koehn et al., 2007). We use a pipeline with the phrase-based decoder with standard parameters, unless noted otherwise. The decoder uses cube pruning (-cube-pruning-pop-limit 2000 -s 2000), MBR (-mbr-size 800 -mbr-scale 1) and monotone at punctuation reordering.

Individual language models (LMs), 5-gram and smoothed using a simplified version of the improved Kneser-Ney method (Chen and Goodman, 1996), are built for each monolingual corpus using IRSTLM 5.80.01 (Federico et al., 2008). These LMs are then interpolated with IRSTLM using the test set of WMT11 as the development set. Finally, the interpolated LMs are merged into one LM preserving the weights using SRILM (Stolcke, 2002).

We use all the parallel corpora available for this language pair: *Europarl* (EU), *News Commentary* (NC), *United Nations* (UN) and *Common Crawl* (CC). Regarding monolingual corpora, we use the freely available monolingual corpora (*Eu-*

[1] http://www.nclt.dcu.ie/mt/

[2] http://www.prompsit.com/

Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 213–218,
Sofia, Bulgaria, August 8-9, 2013 ©2013 Association for Computational Linguistics

roparl, *News Commentary*, *News 2007–2012*) as well as the target side of several parallel corpora: *Common Crawl*, *United Nations* and 10^9 French–English corpus (only for English as target language). Both the parallel and monolingual data are tokenised and truecased using scripts from the MOSES toolkit.

2.2 Data selection

The main contribution in our participation regards the selection of parallel data. We follow the perplexity-based approach to filter monolingual data (Moore and Lewis, 2010) extended to filter parallel data (Axelrod et al., 2011). In our case, we do not measure perplexity only on word forms but also using different types of linguistic information (lemmas and named entities) (Toral, 2013).

We build LMs for the source and target sides of the domain-specific corpus (in our case NC) and for a random subset of the non-domain-specific corpus (EU, UN and CC) of the same size (number of sentences) of the domain-specific corpus. Each parallel sentence s in the non-domain-specific corpus is then scored according to equation 1 where $PP_{Isl}(s)$ is the perplexity of s in the source side according to the domain-specific LM and $PP_{Osl}(s)$ is the perplexity of s in the source side according to the non-domain-specific LM. $PP_{Itl}(s)$ and $PP_{O}tl(s)$ contain the corresponding values for the target side.

$$\begin{aligned} score(s) &= \frac{1}{2} \times (PP_I sl(s) - PP_O sl(s)) \\ &\quad +(PP_I tl(s) - PP_O tl(s)) \qquad (1) \end{aligned}$$

Table 1 shows the results obtained using four models: word forms (*forms*), forms and named entities (*forms+nes*), lemmas (*lem*) and lemmas and named entities (*lem+nes*). Details on these methods can be found in Toral (2013).

For each corpus we selected two subsets (see in bold in Table 1), the one for which one method obtained the best perplexity (top 5% of EU using forms, 2% of UN using lemmas and 50% of CC using forms and named entities) and a bigger one used to compare the performance in SMT (top 14% of EU using lemmas and named entities (*lem+nes*), top 12% of UN using forms and named entities and the whole CC). These subsets are used as training data in our systems.

As we can see in the table, the use of linguistic information allows to obtain subsets with lower perplexity than using solely word forms, e.g. 1057.7 (*lem+nes*) versus 1104.8 (*forms*) for 14% of EU. The only exception to this is the subset that comprises the top 5% of EU, where perplexity using word forms (957.9) is the lowest one.

corpus	size	forms	forms+nes	lem	lem+nes
EU	5%	**957.9**	987.2	974.3	1005.5
	14%	1104.8	1058.7	1111.6	**1057.7**
UN	2%	877.1	969.6	**866.6**	962.2
	12%	1203.2	**1130.9**	1183.8	1131.6
CC	50%	573.0	547.2	574.5	**546.4**
	100%	560.1	560.1	560.1	560.1

Table 1: Perplexities in data selection

2.3 Results

Table 2 presents the results obtained. Note that these were obtained during development and thus the systems are tuned on WMT's 2011 test set and tested on WMT's 2012 test set.

All the systems share the same LM. The first system (*no selection*) is trained with the whole NC and EU. The second (*small*) and third (*big*) systems use as training data the whole NC and subsets of EU (5% and 14%, respectively), UN (2% and 12%, respectively) and CC (50% and 100%, respectively), as shown in Table 1.

System	#sent.	BLEU	BLEUcased
no selection	2.1M	31.99	30.96
small	1.4M	33.12	32.05
big	3.8M	33.49	32.43

Table 2: Number of sentences and BLEU scores obtained on the WMT12 test set for the different systems on the EN–ES translation task.

The advantage of data selection is clear. The second system, although smaller in size compared to the first (1.4M sentence pairs versus 2.1M), takes its training from a more varied set of data, and its performance is over one absolute BLEU point higher.

When comparing the two systems that rely on data selection, one might expect the one that uses data with lower perplexity (*small*) to perform better. However, this is not the case, the third system (*big*) performing around half an absolute BLEU point higher than the second (*small*). This hints at the fact that perplexity alone is not an optimal metric for data selection, but size should also be considered. Note that the size of system 3's phrase table is more than double that of system 2.

3 French-English

This section describe the particularities of the MT systems built for the French-English language pair in both directions. The goal of the experimental setup presented here is to evaluate the gain of adding small in-domain parallel data into a translation system built on a sub-sample of the out-of-domain parallel data.

3.1 Data Pre-processing

All the available parallel and monolingual data for the French-English language pair, including the last versions of *LDC Gigaword* corpora, are normalised and special characters are escaped using the scripts provided by the shared task organisers. Then, the corpora are tokenised and for each language a true-case model is built on the concatenation of all the data after removing duplicated sentences, using the scripts included in MOSES distribution. The corpora are then true-cased before being used to build the language and the translation models.

3.2 Language Model

To build our final language models, we first build LMs on each corpus individually. All the monolingual corpora are considered, as well as the source or target side of the parallel corpora if the data are not already in the monolingual data. We build modified Kneser-Ney discounted 5-gram LMs using the SRILM toolkit for each corpus and separate the LMs in three groups: one in-domain (containing news-commentary and news crawl corpora), another out-of-domain (containing *Common Crawl*, *Europarl*, *UN* and 10^9 corpora), and the last one with *LDC Gigaword* LMs (the data are kept separated by news source, as distributed by *LDC*). The LMs in each group are linearly interpolated based on their perplexities obtained on the concatenation of all the development sets from previous WMT translation tasks. The same development corpus is used to linearly interpolate the in-domain and *LDC* LMs. We finally obtain two LMs, one containing out-of-domain data which is only used to filter parallel data, and another one containing in-domain data which is used to filter parallel data, tuning the translation model weights and at decoding time. Details about the number of n-grams in each language model are presented in Table 3.

	French		English	
	out	in	out	in
1-gram	4.0	3.3	4.2	10.7
2-gram	43.0	44.0	48.2	161.9
3-gram	54.2	61.8	63.4	256.8
4-gram	99.7	119.2	103.2	502.7
5-gram	136.4	165.0	125.4	680.7

Table 3: Number of n-grams (in millions) for the in-domain and out-of-domain LMs in French and English.

3.3 Translation Model

Two phrase-based translation models are built using MGIZA++ (Gao and Vogel, 2008) and MOSES[3], with the default alignment heuristic (*grow-diag-final*) and bidirectional reordering models. The first translation model is in-domain, built with the news-commentary corpus. The second one is built on a sample of all the other parallel corpora available for the French-English language pair. Both corpora are cleaned using the script provided with Moses, keeping the sentences with a length below 80 words. For the second translation model, we used the modified Moore-Lewis method based on the four LMs (two per language) presented in section 3.2. The sum of the source and target perplexity difference is computed for each sentence pair of the corpus. We set an acceptance threshold to keep a limited amount of sentence pairs. The kept sample finally contains $\sim$ 3.7M sentence pairs to train the translation model. Statistics about this data sample and the news-commentary corpus are presented in Table 4. The test set of WMT12 translation task is used to optimise the weights for the two translation models with the MERT algorithm. For this tuning step, the limit of target phrases loaded per source phrase is set to 50. We also use a reordering constraint around punctuation marks. The same parameters are used during the decoding of the test set.

	news-commentary	sample
tokens FR	4.7M	98.6M
tokens EN	4.0M	88.0M
sentences	156.5k	3.7M

Table 4: Statistics about the two parallel corpora, after pre-processing, used to train the translation models.

[3] Moses version 1.0

3.4 Results

The two translation models presented in Section 3.3 allow us to design three translation systems: one using only the in-domain model, one using only the model built on the sub-sample of the out-of-domain data, and one using both models by giving two decoding paths to Moses. For this latter system, the MERT algorithm is also used to optimise the translation model weights. Results obtained on the WMT13 test set, measured with the official automatic metrics, are presented in Table 5. The submitted system is the one built on the sub-sample of the out-of-domain parallel data. This system was chosen during the tuning step because it reached the highest BLEU scores on the development corpus, slightly above the combination of the two translation models.

	News-Com.	Sample	Comb.
	FR-EN		
BLEUdev	26.9	30.0	29.9
BLEU	27.0	30.8	30.4
BLEUcased	26.1	29.8	29.3
TER	62.9	58.9	59.3
	EN-FR		
BLEUdev	27.1	29.7	29.6
BLEU	26.6	29.6	29.4
BLEUcased	25.8	28.7	28.5
TER	65.1	61.8	62.0

Table 5: BLEU and TER scores obtained by our systems. BLEUdev is the score obtained on the development set given by MERT, while BLEU, BLEUcased and TER are obtained on the test set given by the submission website.

For both FR-EN and EN-FR tasks, the best results are reached by the system built on the sub-sample taken from the out-of-domain parallel data. Using only News-Commentary to build a translation model leads to acceptable BLEU scores, with regards to the size of the training corpus. When the sub-sample of the out-of-domain parallel data is used to build the translation model, adding a model built on *News-Commentary* does not improve the results. The difference between these two systems in terms of BLEU score (both cased sensitive and insensitive) indicates that similar results can be achieved, however it appears that the amount of sentence pairs in the sample is large enough to limit the impact of the small in-domain corpus parallel. Further experiments are still required to determine the minimum sample size needed to outperform both the in-domain system and the combination of the two translation models.

4 German-English

In this section we describe our work on German to English subtask. Firstly we describe the Dependency tree to string method which we tried but unfortunately failed due to short of time. Secondly we discuss the baseline system and the preprocessing we performed. Thirdly a system combination method is described.

4.1 Dependency Tree to String Method

Our original plan was to address the long distance reordering problem in German-English translation. We use Xie's Dependency tree to string method(Xie et al., 2011) which obtains good results on Chinese to English translation and exhibits good performance at long distance reordering as our decoder.

We use Stanford dependency parser[4] to parse the English side of the data and Mate-Tool[5] for the German side. The first set of experiments did not lead to encouraging results and due to insufficient time, we decide to switch to other decoders, based on statistical phrase-based and hierarchical approaches.

4.2 Baseline System

In this section we describe the three baseline system we used as well as the preprocessing technologies and the experiments set up.

4.2.1 Preprocessing and Corpus

We first use the normalisation scripts provided by WMT2013 to normalise both English and German side. Then we escape special characters on both sides. We use Stanford tokeniser for English and OpenNLP tokeniser[6] for German. Then we train a true-case model using with *Europarl* and *News-Commentary* corpora, and true-case all the corpus we used. The parallel corpus is filtered with the standard cleaning scripts provided with

[4] `http://nlp.stanford.edu/software/lex-parser.shtml`

[5] http://code.google.com/p/mate-tools/

[6] `http://opennlp.sourceforge.net/models-1.5/`

MOSES. We split the German compound words with jWordSplitter[7].

All the corpus provided for the shared task are used for training our translation models, while WMT2011 and WMT2012 test sets are used to tune the models parameters. For the LM, we use all the monolingual data provided, including *LDC Gigaword*. Each LM is trained with the SRILM toolkit, before interpolating all the LMs according to their weights obtained by minimizing the perplexity on the tuning set (WMT2011 and WMT2012 test sets). As SRILM can only interpolate 10 LMs, we first interpolate a LM with *Europarl*, *News Commentary*, *News Crawl* (2007-2012, each year individually, 6 separate parts), then we interpolate a new LM with this interpolated LM and *LDC Gigawords* (we kept the *Gigaword* subsets separated according to the news sources as distributed by *LDC*, which leads to 7 corpus).

4.2.2 Three baseline systems

We use the data set up described by the former subsection and build up three baseline systems, namely PB MOSES (phrase-based), Hiero MOSES (hierarchical) and CDEC (Dyer et al., 2010). The motivation of choosing Hierarchical Models is to address the German-English's long reorder problem. We want to test the performance of CDEC and Hiero MOSES and choose the best. PB MOSES is used as our benchmark. The three results obtained on the development and test sets for the three baseline system and the system combination are shown in the Table 6.

	Development	Test
PB MOSES	22.0	24.0
Hiero MOSES	22.1	24.4
CDEC	22.5	24.4
Combination	23.0	24.8

Table 6: BLEU scores obtained by our systems on the development and test sets for the German to English translation task.

From the Table 6 we can see that on development set, CDEC performs the best, and its much better than MOSES's two decoder, but on test set, Hiero MOSES and CDEC performs as well as each other, and they both performs better than PB Model.

4.3 System Combination

We also use a word-level combination strategy (Rosti et al., 2007) to combine the three translation hypotheses. To combine these systems, we first use the Minimum Bayes-Risk (MBR) (Kumar and Byrne, 2004) decoder to obtain the 5 best hypothesis as the alignment reference for the Confusion Network (CN) (Mangu et al., 2000). We then use IHMM (He et al., 2008) to choose the backbone build the CN and finally search for and generate the best translation.

We tune the system parameters on development set with Simple-Simplex algorithm. The parameters for system weights are set equal. Other parameters like language model, length penalty and combination coefficient are chosen when we see a good improvement on development set.

5 Conclusion

This paper presented a set of experiments conducted on Spanish-English, French-English and German-English language pairs. For the Spanish-English pair, we have explored the use of linguistic information to select parallel data and use this as the training for SMT. However, the comparison of the performance obtained using this method and the purely statistical one (i.e. perplexity on word forms) remains to be carried out. Another open question regards the optimal size of the selected data. As we have seen, minimum perplexity alone cannot be considered an optimal metric since using a larger set, even if it has higher perplexity, allowed us to obtain notably higher BLEU scores. The question is then how to decide the optimal size of parallel data to select.

For the French-English language pair, we investigated the usefulness of the small in-domain parallel data compared to out-of-domain parallel data sub-sampling. We show that with a sample containing $\sim 3.7M$ sentence pairs extracted from the out-of-domain parallel data, it is not necessary to use the small domain-specific parallel data. Further experiments are still required to determine the minimum sample size needed to outperform both the in-domain system and the combination of the two translation models.

Finally, for the German-English language pair, we presents our exploitation of long ordering problem. We compared two hierarchical models with one phrase-based model, and we also use a system combination strategy to further improve

[7] http://www.danielnaber.de/jwordsplitter/

the translation systems performance.

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement PIAP-GA-2012-324414 (Abu-MaTran) and through Science Foundation Ireland as part of the CNGL (grant 07/CE/I1142).

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 355–362, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Chris Dyer, Jonathan Weese, Hendra Setiawan, Adam Lopez, Ferhan Ture, Vladimir Eidelman, Juri Ganitkevitch, Phil Blunsom, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the ACL 2010 System Demonstrations*, pages 7–12. Association for Computational Linguistics.

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

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57. Association for Computational Linguistics.

Xiaodong He, Mei Yang, Jianfeng Gao, Patrick Nguyen, and Robert Moore. 2008. Indirect-hmm-based hypothesis alignment for combining outputs from machine translation systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 98–107. 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*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Shankar Kumar and William Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *Proceedings of HLT-NAACL*, pages 169–176.

Lidia Mangu, Eric Brill, and Andreas Stolcke. 2000. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech & Language*, 14(4):373–400.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 220–224, Stroudsburg, PA, USA. Association for Computational Linguistics.

Antti-Veikko I Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz, and Bonnie Dorr. 2007. Combining outputs from multiple machine translation systems. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 228–235.

Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. In John H. L. Hansen and Bryan L. Pellom, editors, *INTERSPEECH*. ISCA.

Antonio Toral. 2013. Hybrid Selection of Language Model Training Data Using Linguistic Information and Perplexity. In *Proceedings of the Second Workshop on Hybrid Approaches to Machine Translation (HyTra)*, ACL 2013.

Jun Xie, Haitao Mi, and Qun Liu. 2011. A novel dependency-to-string model for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 216–226. Association for Computational Linguistics.