The CMU-Avenue French-English Translation System

Michael Denkowski **Greg Hanneman** **Alon Lavie**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
{mdenkows,ghannema,alavie}@cs.cmu.edu

Abstract

This paper describes the French-English translation system developed by the Avenue research group at Carnegie Mellon University for the Seventh Workshop on Statistical Machine Translation (NAACL WMT12). We present a method for training data selection, a description of our hierarchical phrase-based translation system, and a discussion of the impact of data size on best practice for system building.

1 Introduction

We describe the French-English translation system constructed by the Avenue research group at Carnegie Mellon University for the shared translation task in the Seventh Workshop on Statistical Machine Translation. The core translation system uses the hierarchical phrase-based model described by Chiang (2007) with sentence-level grammars extracted and scored using the methods described by Lopez (2008). Improved techniques for data selection and monolingual text processing significantly improve the performance of the baseline system.

Over half of all parallel data for the French-English track is provided by the Giga-FrEn corpus (Callison-Burch et al., 2009). Assembled from crawls of bilingual websites, this corpus is known to be noisy, containing sentences that are either not parallel or not natural language. Rather than simply including or excluding the resource in its entirety, we use a relatively simple technique inspired by work in machine translation quality estimation to select the best portions of the corpus for inclusion in our training data. Including around 60% of the Giga-FrEn chosen by this technique yields an improvement of 0.7 BLEU.

Prior to model estimation, we process all parallel and monolingual data using in-house tokenization and normalization scripts that detect word boundaries better than the provided WMT12 scripts. After translation, we apply a monolingual rule-based post-processing step to correct obvious errors and make sentences more acceptable to human judges. The post-processing step alone yields an improvement of 0.3 BLEU to the final system.

We conclude with a discussion of the impact of data size on important decisions for system building. Experimental results show that "best practice" decisions for smaller data sizes do not necessarily carry over to systems built with "WMT-scale" data, and provide some explanation for why this is the case.

2 Training Data

Training data provided for the French-English translation task includes parallel corpora taken from European Parliamentary proceedings (Koehn, 2005), news commentary, and United Nations documents. Together, these sets total approximately 13 million sentences. In addition, a large, web-crawled parallel corpus termed the "Giga-FrEn" (Callison-Burch et al., 2009) is made available. While this corpus contains over 22 million parallel sentences, it is inherently noisy. Many parallel sentences crawled from the web are neither parallel nor sentences. To make use of this large data source, we employ data selection techniques discussed in the next subsection.

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 261–266,
Montréal, Canada, June 7-8, 2012.

Corpus	Sentences
Europarl	1,857,436
News commentary	130,193
UN doc	11,684,454
Giga-FrEn 1stdev	7,535,699
Giga-FrEn 2stdev	5,801,759
Total	27,009,541

Table 1: Parallel training data

Parallel data used to build our final system totals 27 million sentences. Precise figures for the number of sentences in each data set, including selections from the Giga-FrEn, are found in Table 1.

2.1 Data Selection as Quality Estimation

Drawing inspiration from the workshop's featured task, we cast the problem of data selection as one of quality estimation. Specia et al. (2009) report several estimators of translation quality, the most effective of which detect difficult-to-translate source sentences, ungrammatical translations, and translations that align poorly to their source sentences. We can easily adapt several of these predictive features to select good sentence pairs from noisy parallel corpora such as the Giga-FrEn.

We first pre-process the Giga-FrEn by removing lines with invalid Unicode characters, control characters, and insufficient concentrations of Latin characters. We then score each sentence pair in the remaining set (roughly 90% of the original corpus) with the following features:

Source language model: a 4-gram modified Kneser-Ney smoothed language model trained on French Europarl, news commentary, UN doc, and news crawl corpora. This model assigns high scores to grammatical source sentences and lower scores to ungrammatical sentences and non-sentences such as site maps, large lists of names, and blog comments. Scores are normalized by number of n-grams scored per sentence (length + 1). The model is built using the SRILM toolkit (Stolke, 2002).

Target language model: a 4-gram modified Kneser-Ney smoothed language model trained on English Europarl, news commentary, UN doc, and news crawl corpora. This model scores grammaticality on the target side.

Word alignment scores: source-target and target-source MGIZA++ (Gao and Vogel, 2008) force-alignment scores using IBM Model 4 (Och and Ney, 2003). Model parameters are estimated on 2 million words of French-English Europarl and news commentary text. Scores are normalized by the number of alignment links. These features measure the extent to which translations are parallel with their source sentences.

Fraction of aligned words: source-target and target-source ratios of aligned words to total words. These features balance the link-normalized alignment scores.

To determine selection criteria, we use this feature set to score the news test sets from 2008 through 2011 (10K parallel sentences) and calculate the mean and standard deviation of each feature score distribution. We then select two subsets of the Giga-FrEn, "1stdev" and "2stdev". The 1stdev set includes sentence pairs for which the score for *each* feature is above a threshold defined as the development set mean minus one standard deviation. The 2stdev set includes sentence pairs not included in 1stdev that meet the per-feature threshold of mean minus two standard deviations. Hard, per-feature thresholding is motivated by the notion that a sentence pair must meet *all* the criteria discussed above to constitute good translation. For example, high source and target language model scores are irrelevant if the sentences are not parallel.

As primarily news data is used for determining thresholds and building language models, this approach has the added advantage of preferring parallel data in the domain we are interested in translating. Our final translation system uses data from both 1stdev and 2stdev, corresponding to roughly 60% of the Giga-FrEn corpus.

2.2 Monolingual Data

Monolingual English data includes European Parliamentary proceedings (Koehn, 2005), news commentary, United Nations documents, news crawl, the English side of the Giga-FrEn, and the English Gigaword Fourth Edition (Parker et al., 2009). We use all available data subject to the following selection decisions. We apply the initial filter to the Giga-FrEn to remove non-text sections, leaving approximately 90% of the corpus. We exclude the known prob-

Corpus	Words
Europarl	59,659,916
News commentary	5,081,368
UN doc	286,300,902
News crawl	1,109,346,008
Giga-FrEn	481,929,410
Gigaword 4th edition	1,960,921,287
Total	3,903,238,891

Table 2: Monolingual language modeling data (uniqued)

lematic New York Times section of the Gigaword. As many data sets include repeated boilerplate text such as copyright information or browser compatibility notifications, we unique sentences from the UN doc, news crawl, Giga-FrEn, and Gigaword sets by source. Final monolingual data totals 4.7 billion words before uniqueing and 3.9 billion after. Word counts for all data sources are shown in Table 2.

2.3 Text Processing

All monolingual and parallel system data is run through a series of pre-processing steps before construction of the language model or translation model. We first run an in-house normalization script over all text in order to convert certain variably encoded characters to a canonical form. For example, thin spaces and non-breaking spaces are normalized to standard ASCII space characters, various types of "curly" and "straight" quotation marks are standardized as ASCII straight quotes, and common French and English ligatures characters (e.g. œ, fi) are replaced with standard equivalents.

English text is tokenized with the Penn Treebank-style tokenizer attached to the Stanford parser (Klein and Manning, 2003), using most of the default options. We set the tokenizer to Americanize variant spellings such as *color* vs. *colour* or *behavior* vs. *behaviour*. Currency-symbol normalization is avoided.

For French text, we use an in-house tokenization script. Aside from the standard tokenization based on punctuation marks, this step includes French-specific rules for handling apostrophes (French *elision*), hyphens in subject-verb inversions (including the French *t euphonique*), and European-style numbers. When compared to the default WMT12-provided tokenization script, our custom French rules more accurately identify word boundaries, particularly in the case of hyphens. Figure 1 highlights the differences in sample phrases. Subject-verb inversions are broken apart, while other hyphenated words are unaffected; French *aujourd'hui* ("today") is retained as a single token to match English.

Parallel data is run through a further filtering step to remove sentence pairs that, by their length characteristics alone, are very unlikely to be true parallel data. Sentence pairs that contain more than 95 tokens on either side are globally discarded, as are sentence pairs where either side contains a token longer than 25 characters. Remaining pairs are checked for length ratio between French and English, and sentences are discarded if their English translations are either too long or too short given the French length. Allowable ratios are determined from the tokenized training data and are set such that approximately the middle 95% of the data, in terms of length ratio, is kept for each French length.

3 Translation System

Our translation system uses `cdec` (Dyer et al., 2010), an implementation of the hierarchical phrase-based translation model (Chiang, 2007) that uses the KenLM library (Heafield, 2011) for language model inference. The system translates from cased French to cased English; at no point do we lowercase data.

The Parallel data is aligned in both directions using the MGIZA++ (Gao and Vogel, 2008) implementation of IBM Model 4 and symmetrized with the `grow-diag-final` heuristic (Och and Ney, 2003). The aligned corpus is then encoded as a suffix array to facilitate sentence-level grammar extraction and scoring (Lopez, 2008). Grammars are extracted using the heuristics described by Chiang (Chiang, 2007) and feature scores are calculated according to Lopez (2008).

Modified Knesser-Ney smoothed (Chen and Goodman, 1996) n-gram language models are built from the monolingual English data using the SRI language modeling toolkit (Stolke, 2002). We experiment with both 4-gram and 5-gram models.

System parameters are optimized using minimum error rate training (Och, 2003) to maximize the corpus-level cased BLEU score (Papineni et al.,

Base:	`Y a-t-il un collègue pour prendre la parole`
Custom:	`Y a -t-il un collègue pour prendre la parole`
Base:	`Peut-être , à ce sujet , puis-je dire à M. Ribeiro i Castro`
Custom:	`Peut-être , à ce sujet , puis -je dire à M. Ribeiro i Castro`
Base:	`le procès-verbal de la séance d' aujourd' hui`
Custom:	`le procès-verbal de la séance d' aujourd'hui`
Base:	`s' établit environ à 1,2 % du PIB`
Custom:	`s' établit environ à 1.2 % du PIB`

Figure 1: Customized French tokenization rules better identify word boundaries.

pré-éĺectoral	→	pre-electoral
mosaîque	→	mosaique
déragulation	→	deragulation

Figure 2: Examples of cognate translation

2002) on news-test 2008 (2051 sentences). This development set is chosen for its known stability and reliability.

Our baseline translation system uses Viterbi decoding while our final system uses segment-level Minimum Bayes-Risk decoding (Kumar and Byrne, 2004) over 500-best lists using 1 - BLEU as the loss function.

3.1 Post-Processing

Our final system includes a monolingual rule-based post-processing step that corrects obvious translation errors. Examples of correctable errors include capitalization, mismatched punctuation, malformed numbers, and incorrectly split compound words. We finally employ a coarse cognate translation system to handle out-of-vocabulary words. We assume that uncapitalized French source words passed through to the English output are cognates of English words and translate them by removing accents. This frequently leads to (in order of desirability) fully correct translations, correct translations with foreign spellings, or correct translations with misspellings. All of the above are generally preferable to untranslated foreign words. Examples of cognate translations for OOV words in newstest 2011 are shown in Figure 2.[1]

[1]Some OOVs are caused by misspellings in the dev-test source sentences. In these cases we can salvage misspelled English words in place of misspelled French words

	BLEU	(cased)	Meteor	TER
base 5-gram	28.4	27.4	33.7	53.2
base 4-gram	29.1	28.1	34.0	52.5
+1stdev GFE	29.3	28.3	34.2	52.1
+2stdev GFE	29.8	28.9	34.5	51.7
+5g/1K/MBR	29.9	29.0	34.5	51.5
+post-process	30.2	29.2	34.7	51.3

Table 3: Newstest 2011 (dev-test) translation results

4 Experiments

Beginning with a baseline translation system, we incrementally evaluate the contribution of additional data and components. System performance is evaluated on newstest 2011 using BLEU (uncased and cased) (Papineni et al., 2002), Meteor (Denkowski and Lavie, 2011), and TER (Snover et al., 2006). For full consistency with WMT11, we use the NIST scoring script, TER-0.7.25, and Meteor-1.3 to evaluate cased, detokenized translations. Results are shown in Table 3, where each evaluation point is the result of a full tune/test run that includes MERT for parameter optimization.

The baseline translation system is built from 14 million parallel sentences (Europarl, news commentary, and UN doc) and all monolingual data. Grammars are extracted using the "tight" heuristic that requires phrase pairs to be bounded by word alignments. Both 4-gram and 5-gram language models are evaluated. Viterbi decoding is conducted with a cube pruning pop limit (Chiang, 2007) of 200. For this data size, the 4-gram model is shown to significantly outperform the 5-gram.

Adding the 1stdev and 2stdev sets from the Giga-FrEn increases the parallel data size to 27 million

	BLEU	(cased)	Meteor	TER
587M tight	29.1	28.1	34.0	52.5
587M loose	29.3	28.3	34.0	52.5
745M tight	29.8	28.9	34.5	51.7
745M loose	29.6	28.6	34.3	52.0

Table 4: Results for extraction heuristics (dev-test)

	BLEU	(cased)	Meteor	TER
587M 4-gram	29.1	28.1	34.0	52.5
587M 5-gram	28.4	27.4	33.7	53.2
745M 4-gram	29.8	28.9	34.5	51.7
745M 5-gram	29.8	28.9	34.4	51.7

Table 5: Results for language model order (dev-test)

sentences and further improves performance. These runs require new grammars to be extracted, but use the same 4-gram language model and decoding method as the baseline system. With large training data, moving to a 5-gram language model, increasing the cube pruning pop limit to 1000, and using Minimum Bayes-Risk decoding (Kumar and Byrne, 2004) over 500-best lists collectively show a slight improvement. Monolingual post-processing yields further improvement. This decoding/processing scheme corresponds to our final translation system.

4.1 Impact of Data Size

The WMT French-English track provides an opportunity to experiment in a space of data size that is generally not well explored. We examine the impact of data sizes of hundreds of millions of words on two significant system building decisions: grammar extraction and language model estimation. Comparative results are reported on the newstest 2011 set.

In the first case, we compare the "tight" extraction heuristic that requires phrases to be bounded by word alignments to the "loose" heuristic that allows unaligned words at phrase edges. Lopez (2008) shows that for a parallel corpus of 107 million words, using the loose heuristic produces much larger grammars and improves performance by a full BLEU point. However, even our baseline system is trained on substantially more data (587 million words on the English side) and the addition of the Giga-FrEn sets increases data size to 745 million words, seven times that used in the cited work. For each data size, we decode with grammars extracted using each heuristic and a 4-gram language model. As shown in Table 4, the differences are much smaller and the tight heuristic actually produces the best result for the full data scenario. We believe this to be directly linked to word alignment quality: smaller training data results in sparser, noisier word alignments while larger data results in denser, more accurate alignments. In the first case, accumulating unaligned words can make up for shortcomings in alignment quality. In the second, better rules are extracted by trusting the stronger alignment model.

We also compare 4-gram and 5-gram language model performance with systems using tight grammars extracted from 587 million and 745 million sentences. As shown in Table 5, the 4-gram significantly outperforms the 5-gram with smaller data while the two are indistinguishable with larger data[2]. With modified Kneser-Ney smoothing, a lower order model will outperform a higher order model if the higher order model constantly backs off to lower orders. With stronger grammars learned from larger parallel data, the system is able to produce output that matches longer n-grams in the language model.

5 Summary

We have presented the French-English translation system built for the NAACL WMT12 shared translation task, including descriptions of our data selection and text processing techniques. Experimental results have shown incremental improvement for each addition to our baseline system. We have finally discussed the impact of the availability of WMT-scale data on system building decisions and provided comparative experimental results.

References

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proc. of ACL WMT 2009*.

[2]We find that for the full data system, also increasing the cube pruning pop limit and using MBR decoding yields a very slight improvement with the 5-gram model over the same decoding scheme with the 4-gram.

Stanley F. Chen and Joshua Goodman. 1996. An Empirical Study of Smoothing Techniques for Language Modeling. In *Proc. of ACL 1996*.

David Chiang. 2007. Hierarchical Phrase-Based Translation.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proc. of the EMNLP WMT 2011*.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Johnathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A Decoder, Alignment, and Learning Framework for Finite-State and Context-Free Translation Models. In *Proc. of ACL 2010*.

Qin Gao and Stephan Vogel. 2008. Parallel Implementations of Word Alignment Tool. In *Proc. of ACL WSETQANLP 2008*.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proc. of EMNLP WMT 2011*.

Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. In *Proc. of ACL 2003*.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proc. of MT Summit 2005*.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-Risk Decoding for Statistical Machine Translation. In *Proc. of NAACL/HLT 2004*.

Adam Lopez. 2008. Tera-Scale Translation Models via Pattern Matching. In *Proc. of COLING 2008*.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29.

Franz Josef Och. 2003. Minimum Error Rate Training for Statistical Machine Translation. In *Proc. of ACL 2003*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. of ACL 2002*.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2009. English Gigaword Fourth Edition. Linguistic Data Consortium, LDC2009T13.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proc. of AMTA 2006*.

Lucia Specia, Craig Saunders, Marco Turchi, Zhuoran Wang, and John Shawe-Taylor. 2009. Improving the Confidence of Machine Translation Quality Estimates. In *Proc. of MT Summit XII*.

Andreas Stolke. 2002. SRILM - an Extensible Language Modeling Toolkit. In *Proc. of ICSLP 2002*.