An Improved Statistical Transfer System for French–English Machine Translation

Greg Hanneman, Vamshi Ambati, Jonathan H. Clark, Alok Parlikar, Alon Lavie
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213 USA
{ghannema, vamshi, jhclark, aup, alavie}@cs.cmu.edu

Abstract

This paper presents the Carnegie Mellon University statistical transfer MT system submitted to the 2009 WMT shared task in French-to-English translation. We describe a syntax-based approach that incorporates both syntactic and non-syntactic phrase pairs in addition to a syntactic grammar. After reporting development test results, we conduct a preliminary analysis of the coverage and effectiveness of the system's components.

1 Introduction

The statistical transfer machine translation group at Carnegie Mellon University has been developing a hybrid approach combining a traditional rule-based MT system and its linguistically expressive formalism with more modern techniques of statistical data processing and search-based decoding. The Stat-XFER framework (Lavie, 2008) provides a general environment for building new MT systems of this kind. For a given language pair or data condition, the framework depends on two main resources extracted from parallel data: a probabilistic bilingual lexicon, and a grammar of probabilistic synchronous context-free grammar rules. Additional monolingual data, in the form of an n-gram language model in the target language, is also used. The statistical transfer framework operates in two stages. First, the lexicon and grammar are applied to synchronously parse and translate an input sentence; all reordering is applied during this stage, driven by the syntactic grammar. Second, a monotonic decoder runs over the lattice of scored translation pieces produced during parsing and assembles the highest-scoring overall translation according to a log-linear feature model.

Since our submission to last year's Workshop on Machine Translation shared translation task (Hanneman et al., 2008), we have made numerous improvements and extensions to our resource extraction and processing methods, resulting in significantly improved translation scores. In Section 2 of this paper, we trace our current methods for data resource management for the Stat-XFER submission to the 2009 WMT shared French–English translation task. Section 3 explains our tuning procedure, and Section 4 gives our experimental results on various development sets and offers some preliminary analysis.

2 System Construction

Because of the additional data resources provided for the 2009 French–English task, our system this year is trained on nearly eight times as much data as last year's. We used three officially provided data sets to make up a parallel corpus for system training: version 4 of the Europarl corpus (1.43 million sentence pairs), the News Commentary corpus (0.06 million sentence pairs), and the pre-release version of the new Giga-FrEn corpus (8.60 million sentence pairs)[1]. The combined corpus of 10.09 million sentence pairs was preprocessed to remove blank lines, sentences of 80 words or more, and sentence pairs where the ratio between the number of English and French words was larger than 5 to 1 in either direction. These steps removed approximately 3% of the corpus. Given the filtered corpus, our data preparation pipeline proceeded according to the descriptions below.

[1] Because of data processing time, we were unable to use the larger verions 1 or 2 of Giga-FrEn released later in the evaluation period.

Proceedings of the Fourth Workshop on Statistical Machine Translation , pages 140–144,
Athens, Greece, 30 March – 31 March 2009.

2.1 Parsing and Word Alignment

We parsed both sides of our parallel corpus with independent automatic constituency parsers. We used the Berkeley parser (Petrov and Klein, 2007) for both English and French, although we obtained better results for French by tokenizing the data with our own script as a preprocessing step and not allowing the parser to change it. There were approximately 220,000 English sentences that did not return a parse, which further reduced the size of our training corpus by 2%.

After parsing, we re-extracted the leaf nodes of the parse trees and statistically word-aligned the corpus using a multi-threaded implementation (Gao and Vogel, 2008) of the GIZA++ program (Och and Ney, 2003). Unidirectional alignments were symmetrized with the "grow-diag-final" heuristic (Koehn et al., 2005).

2.2 Phrase Extraction and Combination

Phrase extraction for last year's statistical transfer system used automatically generated parse trees on both sides of the corpus as absolute constraints: a syntactic phrase pair was extracted from a given sentence only when a contiguous sequence of English words exactly made up a syntactic constituent in the English parse tree and could also be traced though symmetric word alignments to a constituent in the French parse tree. While this "tree-to-tree" extraction method is precise, it suffers from low recall and results in a low-coverage syntactic phrase table. Our 2009 system uses an extended "tree-to-tree-string" extraction process (Ambati and Lavie, 2008) in which, if no suitable equivalent is found in the French parse tree for an English node, a copy of the English node is projected into the French tree, where it spans the French words aligned to the yield of the English node. This method can result in a 50% increase in the number of extracted syntactic phrase pairs. Each extracted phrase pair retains a syntactic category label; in our current system, the node label in the English parse tree is used as the category for both sides of the bilingual phrase pair, although we subsequently map the full set of labels used by the Berkeley parser down to a more general set of 19 syntactic categories.

We also ran "standard" phrase extraction on the same corpus using Steps 4 and 5 of the Moses statistical machine translation training script (Koehn et al., 2007). The two types of phrases were then merged in a syntax-prioritized combination that removes all Moses-extracted phrase pairs that have source sides already covered by the tree-to-tree-string syntactic phrase extraction. The syntax prioritization has the advantage of still including a selection of non-syntactic phrases while producing a much smaller phrase table than a direct combination of all phrase pairs of both types. Previous experiments we conducted indicated that this comes with only a minor drop in automatic metric scores.

In our current submission, we modify the procedure slightly by removing singleton phrase pairs from the syntactic table before the combination with Moses phrases. The coverage of the combined table is not affected — our syntactic phrase extraction algorithm produces a subset of the non-syntactic phrase pairs extracted from Moses, up to phrase length constraints — but the removal allows Moses-extracted versions of some phrases to survive syntax prioritization. In effect, we are limiting the set of category-labeled syntactic translations we trust to those that have been seen more than once in our training data. For a given syntactic phrase pair, we also remove all but the most frequent syntactic category label for the pair; this removes a small number of entries from our lexicon in order to limit label ambiguity, but does not affect coverage.

From our training data, we extracted 27.6 million unique syntactic phrase pairs after singleton removal, reducing this set to 27.0 million entries after filtering for category label ambiguity. Some 488.7 million unique phrase pairs extracted from Moses were reduced to 424.0 million after syntax prioritization. (The remaining 64.7 million phrase pairs had source sides already covered by the 27.0 million syntactically extracted phrase pairs, so they were thrown out.) This means non-syntactic phrases outnumber syntactic phrases by nearly 16 to 1. However, when filtering the phrase table to a particular development or test set, we find the syntactic phrases play a larger role, as this ratio drops to approximately 3 to 1.

Sample phrase pairs from our system are shown in Figure 1. Each pair includes two rule scores, which we calculate from the source-side syntactic category (c_s), source-side text (w_s), target-side category (c_t), and target-side text (w_t). In the case of Moses-extracted phrase pairs, we use the "dummy" syntactic category PHR. Rule score $r_{t|s}$ is a maximum likelihood estimate of the distri-

c_s	c_t	w_s	w_t	$r_{t\|s}$	$r_{s\|t}$
ADJ	ADJ	espagnols	Spanish	0.8278	0.1141
N	N	représentants	officials	0.0653	0.1919
NP	NP	représentants de la Commission	Commission officials	0.0312	0.0345
PHR	PHR	haute importance à	very important to	0.0357	0.0008
PHR	PHR	est chargé de	has responsibility for	0.0094	0.0760

Figure 1: Sample lexical entries, including non-syntactic phrases, with rule scores (Equations 1 and 2).

bution of target-language translations and source- and target-language syntactic categories given the source string (Equation 1). The $r_{s|t}$ score is similar, but calculated in the reverse direction to give a source-given-target probability (Equation 2).

$$r_{t|s} = \frac{\#(w_t, c_t, w_s, c_s)}{\#(w_s) + 1} \quad (1)$$

$$r_{s|t} = \frac{\#(w_t, c_t, w_s, c_s)}{\#(w_t) + 1} \quad (2)$$

Add-one smoothing in the denominators counteracts overestimation of the rule scores of lexical entries with very infrequent source or target sides.

2.3 Syntactic Grammar

Syntactic phrase extraction specifies a node-to-node alignment across parallel parse trees. If these aligned nodes are used as decomposition points, a set of synchronous context-free rules that produced the trees can be collected. This is our process of syntactic grammar extraction (Lavie et al., 2008). For our 2009 WMT submission, we extracted 11.0 million unique grammar rules, 9.1 million of which were singletons, from our parallel parsed corpus. These rules operate on our syntactically extracted phrase pairs, which have category labels, but they may also be partially lexicalized with explicit source or target word strings. Each extracted grammar rule is scored according to Equations 1 and 2, where now the right-hand sides of the rule are used as w_s and w_t.

As yet, we have made only minimal use of the Stat-XFER framework's grammar capabilities, especially for large-scale MT systems. For the current submission, the syntactic grammar consisted of 26 manually chosen high-frequency grammar rules that carry out some reordering between English and French. Since rules for high-level reordering (near the top of the parse tree) are unlikely to be useful unless a large amount of parse structure can first be built, we concentrate our rules on low-level reorderings taking place within or around small constituents. Our focus for this selection is the well-known repositioning of adjectives and adjective phrases when translating from French to English, such as from *le Parlement européen* to *the European Parliament* or from *l' intervention forte et substantielle* to *the strong and substantial intervention*. Our grammar thus consists of 23 rules for building noun phrases, two rules for building adjective phrases, and one rule for building verb phrases.

2.4 English Language Model

We built a suffix-array language model (Zhang and Vogel, 2006) on approximately 700 million words of monolingual data: the unfiltered English side of our parallel training corpus, plus the 438 million words of English monolingual news data provided for the WMT 2009 shared task. With the relatively large amount of data available, we made the somewhat unusual decision of building our language model (and all other data resources for our system) in mixed case, which adds approximately 12.3% to our vocabulary size. This saves us the need to build and run a recaser as a postprocessing step on our output. Our mixed-case decision may also be validated by preliminary test set results, which show that our submission has the smallest drop in BLEU score (0.0074) between uncased and cased evaluation of any system in the French–English translation task.

3 System Tuning

Stat-XFER uses a log-linear combination of seven features in its scoring of translation fragments: language model probability, source-given-target and target-given-source rule probabilities, source-given-target and target-given-source lexical probabilities, a length score, and a fragmentation score based on the number of parsed translation fragments that make up the output sentence. We tune the weights for these features with several rounds of minimum error rate training, optimizing to-

Data Set	Primary			Contrastive		
	METEOR	BLEU	TER	METEOR	BLEU	TER
news-dev2009a-425	0.5437	0.2299	60.45	—	—	—
news-dev2009a-600	—	—	—	0.5134	0.2055	63.46
news-dev2009b	0.5263	0.2073	61.96	0.5303	0.2104	61.74
nc-test2007	0.6194	0.3282	51.17	0.6195	0.3226	51.49

Figure 2: Primary and contrastive system results on tuning and development test sets.

wards the BLEU metric. For each tuning iteration, we save the n-best lists output by the system from previous iterations and concatenate them onto the current n-best list in order to present the optimizer with a larger variety of translation outputs and score values.

From the provided "news-dev2009a" development set we create two tuning sets: one using the first 600 sentences of the data, and a second using the remaining 425 sentences. We tuned our system separately on each set, saving the additional "news-dev2009b" set as a final development test to choose our primary and contrastive submissions[2]. At run time, our full system takes on average between four and seven seconds to translate each input sentence, depending on the size of the final bilingual lexicon.

4 Evaluation and Analysis

Figure 2 shows the results of our primary and contrastive systems on four data sets. First, we report final (tuned) performance on our two tuning sets — the last 425 sentences of news-dev2009a for the primary system, and the first 600 sentences of the same set for the contrastive. We also include our development test (news-dev2009b) and, for additional comparison, the "nc-test2007" news commentary test set from the 2007 WMT shared task. For each, we give case-insensitive scores on version 0.6 of METEOR (Lavie and Agarwal, 2007) with all modules enabled, version 1.04 of IBM-style BLEU (Papineni et al., 2002), and version 5 of TER (Snover et al., 2006).

From these results, we highlight two interesting areas of analysis. First, the low tuning and development test set scores bring up questions about system coverage, given that the news domain was not strongly represented in our system's training data. We indeed find a significantly larger proportion of out-of-vocabulary (OOV) words in news-domain sets: the news-dev2009b set is translated by our primary submission with 402 of 6263 word types (6.42%) or 601 of 27,821 word tokens (2.16%) unknown. The same system running on the 2007 WMT "test2007" set of Europarl-derived data records an OOV rate of only 87 of 7514 word types (1.16%) or 105 of 63,741 word tokens (0.16%).

Second, we turn our attention to the usefulness of the syntactic grammar. Though small, we find it to be both beneficial and precise. In the 1026-sentence news-dev2009b set, for example, we find 351 rule applications — the vast majority of them (337) building noun phrases. The three most frequently occurring rules are those for reordering the sequence [DET N ADJ] to [DET ADJ N] (52 occurrences), the sequence [N ADJ] to [ADJ N] (51 occurrences), and the sequence [N^1 *de* N^2] to [N^2 N^1] (45 occurrences). We checked precision by manually reviewing the 52 rule applications in the first 150 sentences of news-dev2009b. There, 41 of the occurrences (79%) were judged to be correct and beneficial to translation output. Of the remainder, seven were judged incorrect or detrimental and four were judged either neutral or of unclear benefit.

We expect to continue to analyze the output and effectiveness of our system in the coming months. In particular, we would like to learn more about the usefulness of our 26-rule grammar with the view of using significantly larger grammars in future versions of our system.

Acknowledgments

This research was supported in part by NSF grants IIS-0121631 (AVENUE) and IIS-0534217 (LETRAS), and by the DARPA GALE program. We thank Yahoo! for the use of the M45 research computing cluster, where we ran the parsing stage of our data processing.

[2]Due to a data processing error, the choice of the primary submission was based on incorrectly computed scores. In fact, the contrastive system has better performance on our development test set.

References

Vamshi Ambati and Alon Lavie. 2008. Improving syntax driven translation models by re-structuring divergent and non-isomorphic parse tree structures. In *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*, pages 235–244, Waikiki, HI, October.

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, Columbus, OH, June.

Greg Hanneman, Edmund Huber, Abhaya Agarwal, Vamshi Ambati, Alok Parlikar, Erik Peterson, and Alon Lavie. 2008. Statistical transfer systems for French–English and German–English machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 163–166, Columbus, OH, June.

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*, Pittsburgh, PA, October.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic, June.

Alon Lavie, Alok Parlikar, and Vamshi Ambati. 2008. Syntax-driven learning of sub-sentential translation equivalents and translation rules from parsed parallel corpora. In *Proceedings of the Second ACL Workshop on Syntax and Structure in Statistical Translation*, pages 87–95, Columbus, OH, June.

Alon Lavie. 2008. Stat-XFER: A general search-based syntax-driven framework for machine translation. In *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, pages 362–375. Springer.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evalution of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, July.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Proceedings of NAACL HLT 2007*, pages 404–411, Rochester, NY, April.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the Seventh Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, MA, August.

Ying Zhang and Stephan Vogel. 2006. Suffix array and its applications in empirical natural language processing. Technical Report CMU-LTI-06-010, Carnegie Mellon University, Pittsburgh, PA, December.