Statistical Transfer Systems for French–English and German–English Machine Translation

Greg Hanneman and **Edmund Huber** and **Abhaya Agarwal** and **Vamshi Ambati** and **Alok Parlikar** and **Erik Peterson** and **Alon Lavie**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213 USA
{ghannema, ehuber, abhayaa, vamshi, aup, eepeter, alavie}@cs.cmu.edu

Abstract

We apply the Stat-XFER statistical transfer machine translation framework to the task of translating from French and German into English. We introduce statistical methods within our framework that allow for the principled extraction of syntax-based transfer rules from parallel corpora given word alignments and constituency parses. Performance is evaluated on test sets from the 2007 WMT shared task.

1 Introduction

The Carnegie Mellon University statistical transfer (Stat-XFER) framework is a general search-based and syntax-driven framework for developing MT systems under a variety of data conditions (Lavie, 2008). At its core is a transfer engine using two language-pair-dependent resources: a grammar of weighted synchronous context-free rules (possibly augmented with unification-style feature constraints), and a probabilistic bilingual lexicon of syntax-based word- and phrase-level translations. The Stat-XFER framework has been used to develop research MT systems for a number of language pairs, including Chinese–English, Hebrew–English, Urdu–English, and Hindi–English.

In this paper, we describe our use of the framework to create new French–English and German–English MT systems for the 2008 Workshop on Statistical Machine Translation shared translation task. We first describe the acquisition and processing of resources for each language pair and the roles of those resources within the Stat-XFER system (Section 2); we then report results on common test sets (Section 3) and share some early analysis and future directions (Section 4).

2 System Description

Building a new machine translation system under the Stat-XFER framework involves constructing a bilingual translation lexicon and a transfer grammar. Over the past six months, we have developed new methods for extracting syntax-based translation lexicons and transfer rules fully automatically from parsed and word-aligned parallel corpora. These new methods are described in detail by Lavie et al. (2008). Below, we detail the statistical methods by which these resources were extracted for our French–English and German–English systems.

2.1 Lexicon

The bilingual lexicon is automatically extracted from automatically parsed and word-aligned parallel corpora. To obtain high-quality statistical word alignments, we run GIZA++ (Och and Ney, 2003) in both the source-to-target and target-to-source directions, then combine the resulting alignments with the Sym2 symmetric alignment heuristic of Ortiz-Martínez et al. (2005)[1]. From this data, we extract a lexicon of both word-to-word and syntactic phrase-to-phrase translation equivalents.

The word-level correspondences are extracted directly from the word alignments: parts of speech for these lexical entries are obtained from the preter-

[1] We use Sym2 over more well-known heuristics such as "grow-diag-final" because Sym2 has been shown to give the best results for the node-alignment subtask that is part of our processing chain.

Proceedings of the Third Workshop on Statistical Machine Translation, pages 163–166,
Columbus, June 2008.

w_s	c_s	w_t	c_t	r
paru	V	appeared	V	0.2054
paru	V	seemed	V	0.1429
paru	V	found	V	0.0893
paru	V	published	V	0.0804
paru	V	felt	V	0.0714
⋮		⋮		⋮
paru	V	already	ADV	0.0089
paru	V	appear	V	0.0089
paru	V	authoritative	ADJ	0.0089

Table 1: Part of the lexical entry distribution for the French (source) word *paru*.

minal nodes of parse trees of the source and target sentences. If parsers are unavailable for either language, we have also experimented with determining parts of speech with independent taggers such as TreeTagger (Schmid, 1995). Alternatively, parts of speech may be projected through the word alignments from one language to the other if the information is available on at least one side. Syntactic phrase-level correspondences are extracted from the parallel data by applying the PFA node alignment algorithm described by Lavie et al. (2008). The yields of the aligned parse tree nodes are extracted as constituent-level translation equivalents.

Each entry in the lexicon is assigned a rule score, r, based on its source-side part of speech c_s, source-side text w_s, target-side part of speech c_t, and target-side text w_t. The score is a maximum-likelihood estimate of the distribution of target-language translation and source- and target-language parts of speech, given the source word or phrase.

$$
\begin{aligned}
r &= p(w_t, c_t, c_s \mid w_s) && (1) \\
&\approx \frac{\#(w_t, c_t, w_s, c_s)}{\#(w_s) + 1} && (2)
\end{aligned}
$$

We employ add-one smoothing in the denominator of Equation 2 to counteract overestimation in the case that $\#(w_s)$ is small. Rule scores provide a way to promote the more likely translation alternatives while still retaining a high degree of diversity in the lexicon. Table 1 shows part of the lexical distribution for the French (source) word *paru*.

The result of the statistical word alignment process and lexical extraction is a bilingual lexicon containing 1,064,755 entries for French–English and 1,111,510 entries for German–English. Sample lexical entries are shown in Figure 1.

2.2 Grammar

Transfer grammars for our earlier statistical transfer systems were manually created by in-house experts of the languages involved and were therefore small. The Stat-XFER framework has since been extended with procedures for automatic grammar acquisition from a parallel corpus, given constituency parses for source or target data or both. Our French and German systems used the context-free grammar rule extraction process described by Lavie et al. (2008). For French, we used 300,000 parallel sentences from the Europarl training data parsed on the English side with the Stanford parser (Klein and Manning, 2003) and on the French side with the Xerox XIP parser (Aït-Mokhtar et al., 2001). For German, we used 300,000 Europarl sentence pairs parsed with the English and German versions of the Stanford parser[2].

The set of rules extracted from the parsed corpora was filtered down after scoring to improve system performance and run time. The final French rule set was comprised of the 1500 most frequently occurring rules. For German, rules that occurred less than twice were filtered out, leaving a total of 16,469. In each system, rule scores were again calculated by Equation 2, with w_s and w_t representing the full right-hand sides of the source and target grammar rules.

A secondary version of our French system used a word-level lexicon extracted from the intersection, rather than the symmetricization, of the GIZA++ alignments in each direction; we hypothesize that this tends to improve precision at the expense of recall. The word-level lexicon was supplemented with syntax-based phrase-level entries obtained from the PFA node alignment algorithm. The grammar contained the 700 highest-frequency and the 500 highest-scoring rules extracted from the parallel parsed corpus. This version had a total lexicon size of 2,023,531 entries and a total grammar of 1034 rules after duplicates were removed. Figure 2 shows

[2]Due to a combination of time constraints and paucity of computational resources, only a portion of the Europarl parallel corpus was utilized, and none of the supplementary news commentary training data was integrated.

```
{VS,248840}                              {NP,2000012}
V::V |: ["paru"] -> ["appeared"]         NP::NP |: ["ein" "Beispiel"] -> ["an" "example"]
(                                        (
  (*score* 0.205357142857143)              (*score* 0.763636363636364)
)                                        )
```

Figure 1: Sample lexical entries for French and German.

sample grammar rules automatically learned by the process described above.

2.3 Transfer Engine

The Stat-XFER transfer engine runs in a two-stage process, first applying the grammar and lexicon to an input sentence, then running a decoder over the resulting lattice of scored translation pieces. Scores for each translation piece are based on a log-linear combination of several features: language model probability, rule scores, source-given-target and target-given-source lexical probabilities, parse fragmentation, and length. For more details, see Lavie (2008). The use of a German transfer grammar an order of magnitude larger than the corresponding French grammar was made possible due to a recent optimization made in the engine. When enabled, it constrains the search of translation hypotheses to the space of hypotheses whose structure satisfies the consituent structure of a source-side parse.

3 Evaluation

We trained our model parameters on a subset of the provided "dev2006" development set, optimizing for case-insensitive IBM-style BLEU (Papineni et al., 2002) with several iterations of minimum error rate training on n-best lists. In each iteration's list, we also included the lists from previous iterations in order to maintain a diversity of hypothesis types and scores. The provided "test2007" and "nc-test2007" data sets, identical with the test data from the 2007 Workshop on Statistical Machine Translation shared task, were used as internal development tests.

Tables 2, 3, and 4 report scores on these data sets for our primary French, secondary French, and German systems. We report case-insensitive scores for 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).

Data Set	METEOR	BLEU	TER
dev2006	0.5332	0.2063	64.81
test2007	0.5358	0.2078	64.75
nc-test2007	0.5369	0.1719	69.83

Table 2: Results for the primary French–English system on provided development and development test sets.

Data Set	METEOR	BLEU	TER
dev2006	0.5330	0.2086	65.02
test2007	0.5386	0.2129	64.29
nc-test2007	0.5311	0.1680	70.90

Table 3: Results for the secondary French–English system on provided development and development test sets.

4 Analysis and Conclusions

From the development test results in Section 3, we note that the Stat-XFER systems' performance currently lags behind the state-of-the-art scores on the 2007 test data[3]. This may be in part due to the low volume of training data used for rule learning. A key research question in our approach is how to distinguish low-frequency correct and useful transfer rules from "noisy" rules that are due to parser errors and incorrect word alignments. We believe that learning rules from more data will help alleviate this problem by proportionally increasing the counts of good rules compared to incorrect ones. We also plan to study methods for more effective rule set pruning, regardless of the volume of training data used.

The difference in metric scores between in-domain and out-of-domain data is partly due to effects of reference length on the metrics used. Detailed output from METEOR and BLEU shows that the reference translations for the test2007 set are about 94% as long as the primary French–English

[3]Top scores on the 2007 test data are approximately 0.60 METEOR, 0.33 BLEU, and 57.6 TER. See Callison-Burch et al. (2007) for full results.

```
{PP,1627955}
PP:PP [PRE "d'" "autres" N] -> [PRE "other" N]
(
  (*score* 0.866050808314088 )
  (X1::Y1)
  (X4::Y3)
)
```

```
{PP,3000085}
PP:ADVP ["vor" CARD "Monaten"] -> [NUM "months" "ago"]
(
  (*score* 0.9375)
  (X2::Y1)
)
```

Figure 2: Sample grammar rules for French and German.

Data Set	METEOR	BLEU	TER
dev2006	0.4967	0.1794	68.68
test2007	0.5052	0.1878	67.94
nc-test2007	0.4939	0.1347	74.38

Table 4: Results for the German–English system on provided development and development test sets.

system's translations. On this set, our system has approximately balanced precision (0.62) and recall (0.66). However, the nc-test2007 references are only 84% as long as our output, a situation that hurts our system's precision (0.57) but boosts its recall (0.68). METEOR, as a metric that favors recall, shows a negligible increase in score between these two test sets, while BLEU and TER report significant relative drops of 17.3% and 7.8%. This behavior appears to be consistent on the test2007 and nc-test2007 data sets across systems (Callison-Burch et al., 2007).

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 the members of the Parsing and Semantics group at Xerox Research Centre Europe for assisting in parsing the French data using their XIP parser.

References

Salah Aït-Mokhtar, Jean-Pierre Chanod, and Claude Roux. 2001. A multi-input dependency parser. In *Proceedings of the Seventh International Workshop on Parsing Technologies*, Beijing, China, October.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic, June.

Dan Klein and Christopher D. Manning. 2003. Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems 15*, pages 3–10. MIT Press, Cambridge, MA.

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 Workshop on Syntax and Structure in Statistical Translation*, Columbus, OH, June. To appear.

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.

Daniel Ortiz-Martínez, Ismael García-Varea, and Francisco Casacuberta. 2005. Thot: A toolkit to train phrase-based models for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, pages 141–148, Phuket, Thailand, September.

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.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT Workshop*.

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.