Stanford University's Submissions to the WMT 2014 Translation Task

Julia Neidert[*], **Sebastian Schuster**[*], **Spence Green**,
Kenneth Heafield, and **Christopher D. Manning**
Computer Science Department, Stanford University
{jneid,sebschu,spenceg,heafield,manning}@cs.stanford.edu

Abstract

We describe Stanford's participation in the French-English and English-German tracks of the 2014 Workshop on Statistical Machine Translation (WMT). Our systems used large feature sets, word classes, and an optional unconstrained language model. Among constrained systems, ours performed the best according to uncased BLEU: 36.0% for French-English and 20.9% for English-German.

1 Introduction

Phrasal (Green et al., 2014b) is a phrase-based machine translation system (Och and Ney, 2004) with an online, adaptive tuning algorithm (Green et al., 2013c) which allows efficient tuning of feature-rich translation models. We improved upon the basic Phrasal system with sparse features over word classes, class-based language models, and a web-scale language model.

We submitted one constrained French-English (Fr-En) system, one unconstrained English-German (En-De) system with a huge language model, and one constrained English-German system without it. Each system was built using over 100,000 features and was tuned on over 10,000 sentences. This paper describes our submitted systems and discusses how the improvements affect translation quality.

2 Data Preparation & Post-Processing

We used all relevant data allowed by the constrained condition, with the exception of HindEnCorp and Wiki Headlines, which we deemed too noisy. Specifically, our parallel data consists of the Europarl version 7 (Koehn, 2005), parallel CommonCrawl (Smith et al., 2013), French-English UN, Giga-FrEn, and News Commentary corpora provided by the evaluation. For monolingual data, we used the provided news crawl data from all years, English Gigaword version 5 (Parker et al., 2011), and target sides of the parallel data. This includes English from the Yandex, CzEng, and parallel CommonCrawl corpora. For parallel CommonCrawl, we concatenated the English halves for various language pairs and then deduplicated at the sentence level.

	Sentences	Tokens
En-De	4.5M	222M
Fr-En	36.3M	2.1B

Table 1: Gross parallel corpus statistics after pre-processing.

	Constrained LM	Unconstrained LM
German	1.7B	38.9 B
English	7.2B	-

Table 2: Number of tokens in pre-processed monolingual corpora used to estimate the language models. We split the constrained English data into two models: 3.7 billion tokens from Gigaword and 3.5 billion tokens from all other sources.

In addition, our unconstrained English-German system used German text extracted from the entire 2012, 2013, and winter 2013 CommonCrawl[1] corpora by Buck et al. (2014).

Tables 1 and 2 show the sizes of the pre-processed corpora of parallel text and monolingual text from which our systems were built.

2.1 Pre-Processing

We used Stanford CoreNLP to tokenize the English and German data according to the Penn Treebank standard (Marcus et al., 1993). The French source data was tokenized similarly to the French Treebank

[*] These authors contributed equally.

[1] http://commoncrawl.org

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 150–156,
Baltimore, Maryland USA, June 26–27, 2014. ©2014 Association for Computational Linguistics

(Abeillé et al., 2003) using the Stanford French tokenizer (Green et al., 2013b).

We also lowercased the data and removed any control characters. Further, we filtered out all lines that consisted mainly of punctuation marks, removed characters that are frequently used as bullet points and standardized white spaces and newlines. We additionally filtered out sentences longer than 100 tokens from the parallel corpora in order to speed up model learning.

2.2 Alignment

For both systems, we used the Berkeley Aligner (Liang et al., 2006) with default settings to align the parallel data. We symmetrized the alignments using the grow-diag heuristic.

2.3 Language Models

Our systems used up to three language models.

2.3.1 Constrained Language Models

For En-De, we used lmplz (Heafield et al., 2013) to estimate a 5-gram language model on all WMT German monolingual data and the German side of the parallel Common Crawl corpus. To query the model, we used KenLM (Heafield, 2011).

For the Fr-En system, we also estimated a 5-gram language model from all the monolingual English data and the English side of the parallel Common Crawl, UN, Giga-FrEn, CzEng and Yandex corpora using the same procedure as above. Additionally, we estimated a second language model from the English Gigaword corpus.

All of these language models used interpolated modified Kneser-Ney smoothing (Kneser and Ney, 1995; Chen and Goodman, 1998).

2.3.2 Unconstrained Language Model

Our unconstrained En-De submission used an additional language model trained on German web text gathered by the Common Crawl Foundation and processed by Buck et al. (2014). This corpus was formed from the 2012, 2013, and winter 2013 Common Crawl releases, which consist of web pages converted to UTF-8 encoding with HTML stripped. Applying the Compact Language Detector 2,[2] 2.89% of the data was identified as German, amounting to 1 TB of uncompressed text. After splitting sentences with the Europarl sentence splitter (Koehn, 2005), the text was deduplicated at the sentence level to reduce the impact of boilerplate and pages that appeared in multiple crawls, discarding 78% of the data. We treated the resulting data as normal text, pre-processing it as described in Section 2.1 to yield 38.9 billion tokens. We built an unpruned interpolated modified Kneser-Ney language model with this corpus (Table 3) and added it as an additional feature alongside the constrained language models. At 38.9 billion tokens after deduplication, this monolingual data is almost 23 times as large as the rest of the German monolingual corpus. Since the test data was also collected from the web, we cannot be sure that the test sentences were not in the language model. However, substantial portions of the test set are translations from other languages, which were not posted online until after 2013.

Order	1	2	3	4	5
Count	226	1,916	6,883	13,292	17,576

Table 3: Number of unique n-grams, in millions, appearing in the Common Crawl German language model.

2.3.3 Word-Class Language Model

We also built a word-class language model for the En-De system. We trained 512 word classes on the constrained German data using the predictive one-sided class model of Whittaker and Woodland (2001) with the parallelized clustering algorithm of Uszkoreit and Brants (2008) by Green et al. (2014a). All tokens were mapped to their word class; infrequent tokens appearing fewer than 5 times were mapped to a special cluster for unknown tokens. Finally, we estimated a 7-gram language model on the mapped corpus with SRILM (Stolcke, 2002) using Witten-Bell smoothing (Bell et al., 1990).

2.4 Tuning and Test Data

For development, we tuned our systems on all 13,573 sentences contained in the newstest2008-2012 data sets and tested on the 3,000 sentences of the newstest2013 data set. The final system weights were chosen among all tuning iterations using performance on the newstest2013 data set.

2.5 Post-Processing

Our post-processor recases and detokenizes system output. For the English-German system, we combined both tasks by using a Conditional Random Field (CRF) model (Lafferty et al., 2001) to

[2] https://code.google.com/p/cld2/

learn transformations between the raw output characters and the post-processed versions. For each test dataset, we trained a separate model on 500,000 sentences selected using the Feature Decay Algorithm for bitext selection (Biçici and Yuret, 2011). Features used include the character type of the current and surrounding characters, the token type of the current and surrounding tokens, and the position of the character within its token.

The English output was recased using a language model based recaser (Lita et al., 2003). The language model was trained on the English side of the Fr-En parallel data using lmplz.

3 Translation System

We built our translation systems using Phrasal.

3.1 Features

Our translation model has 19 dense features that were computed for all translation hypotheses: the nine Moses (Koehn et al., 2007) baseline features, the eight hierarchical lexicalized reordering model features by Galley and Manning (2008), the log count of each rule, and an indicator for unique rules. On top of that, the model uses the following additional features of Green et al. (2014a).

Rule indicator features: An indicator feature for each translation rule. To combat overfitting, this feature fires only for rules that occur more than 50 times in the parallel data. Additional indicator features were constructed by mapping the words in each rule to their corresponding word classes.

Target unigram class: An indicator feature for the class of each target word.

Alignments: An indicator feature for each alignment in a translation rule, including multi-word alignments. Again, class-based translation rules were used to extract additional indicator features.

Source class deletion: An indicator feature for the class of each unaligned source word in a translation rule.

Punctuation count ratio: The ratio of target punctuation tokens to source punctuation tokens for each derivation.

Function word ratio: The ratio of target function words to source function words. The function words for each language are the 35 most frequent words on each side of the parallel data. Numbers and punctuation marks are not included in this list.

Target-class bigram boundary: An indicator feature for the concatenation of the word class of the rightmost word in the left rule and the word class of the leftmost word in the right rule in each adjacent rule pair in a derivation.

Length features: Indicator features for the length of the source side and for the length of the target side of the translation rule and an indicator feature for the concatenation of the two lengths.

Rule orientation features: An indicator feature for each translation rule combined with its orientation class (monotone, swap, or discontinuous). This feature also fires only for rules that occur more than 50 times in the parallel data. Again, class-based translation rules were used to extract additional features.

Signed linear distortion: The signed linear distortion δ for two rules a and b is $\delta = r(a) - l(b) + 1$, where $r(x)$ is the rightmost source index of rule x and $l(x)$ is the leftmost source index of rule x. Each adjacent rule pair in a derivation has an indicator feature for the signed linear distortion of this pair.

Many of these features consider word classes instead of the actual tokens. For the target side, we used the same word classes as we used to train the class-based language model. For the source side, we trained word classes on all available data using the same method.

3.2 Tuning

We used an online, adaptive tuning algorithm (Green et al., 2013c) to learn the feature weights. The loss function is an online variant of expected BLEU (Green et al., 2014a). As a sentence-level metric, we used the extended BLEU+1 metric that smooths the unigram precision as well as the reference length (Nakov et al., 2012). For feature selection, we used L_1 regularization. Each tuning epoch produces a different set of weights; we tried all of them on newstest2013, which was held out from the tuning set, then picked the weights that produced the best uncased BLEU score.

3.3 System Parameters

We started off with the parameters of our systems for the WMT 2013 Translation Task (Green et al., 2013a) and optimized the L_1-regularization strength. Both systems used the following tuning parameters: a 200-best list, a learning rate of 0.02 and a mini-batch size of 20. The En-De system

Track	Stanford	Best	Rank
En-De constrained	19.9	20.1	3
En-De unconstrained	20.0	20.6	5
Fr-En constrained	34.5	35.0	3

(a) cased BLEU (%)

Track	Stanford	Best	Rank
En-De constrained	20.7	20.7	1
En-De unconstrained	20.9	21.0	3
Fr-En constrained	36.0	36.0	1

(b) uncased BLEU (%)

Table 4: Official results in terms of cased and uncased BLEU of our submitted systems compared to the best systems for each track. The ranks for the unconstrained system are calculated relative to all primary submissions for the language pair, whereas the ranks for the constrained systems are relative to only the constrained systems submitted.

used a phrase length limit of 8, a distortion limit of 6 and a L_1-regularization strength of 0.0002. The Fr-En system used a phrase length limit of 9, a distortion limit of 5 and a L_1-regularization strength of 0.0001.

During tuning, we set the stack size for cube pruning to Phrasal's default value of 1200. To decode the test set, we increased the stack size to 3000.

4 Results

Table 4 shows the official results of our systems compared to other submissions to the WMT shared task. Both our En-De and Fr-En systems achieved the highest uncased BLEU scores among all constrained submissions. However, our recaser evidently performed quite poorly compared to other systems, so our constrained systems ranked third by cased BLEU score. Our unconstrained En-De submission ranked third among all systems by uncased BLEU and fifth by cased BLEU.

To demonstrate the effectiveness of the individual improvements, we show results for four different En-De systems: (1) A baseline that contains only the 19 dense features, (2) a feature-rich translation system with the additional rich features, (3) a feature-rich translation system with an additional word class LM, and (4) a feature-rich translation system with an additional wordclass LM and a huge language model. For Fr-En we only built systems (1)-(3). Results for all systems can be seen in Table 5 and Table 6. From these results, we can see that both language pairs benefitted from adding rich features (+0.4 BLEU for En-De and +0.5 BLEU for Fr-En). However, we only see improvements from the class-based language model in the case of the En-De system (+0.4 BLEU). For this reason our Fr-En submission did not use a class-based language model. Using additional data in the form of a huge language model further improved our En-De system by almost 1% BLEU on the newstest2013 data set. However, we only saw 0.2 BLEU improvement on the newstest2014 data set.

4.1 Analysis

Gains from rich features are in line with the gains we saw in the WMT 2013 translation task (Green et al., 2013a). We suspect that rich features would improve the translation quality a lot more if we had several reference translations to tune on.

The word class language model seemed to improve only translations in our En-De system while it had no effect on BLEU in our Fr-En system. One of the main reasons seems to be that the 7-gram word class language model helped particularly with long range reordering, which happens far more frequently in the En-De language pair compared to the Fr-En pair. For example, in the following translation, we can see that the system with the class-based language model successfully translated the verb in the second clause (set in *italic*) while the system without the class-based language model did not translate the verb.

Source: It became clear to me that this *is* my path.

Feature-rich: Es wurde mir klar, dass das mein Weg.

Word class LM: Es wurde mir klar, dass das mein Weg *ist*.

We can also see that the long range of the word class language model improved grammaticality as shown in the following example:

Source: Meanwhile, more than 40 percent of the population *are* HIV positive.

Feature-rich: Inzwischen *sind* mehr als 40 Prozent der Bevölkerung *sind* HIV positiv.

	#iterations	tune	2013	2013 cased	2014	2014 cased
Dense	8	16.9	19.6	18.7	20.0	19.2
Feature-rich	10	20.1	20.0	19.0	20.0	19.2
+ Word class LM	15	21.1	20.4	19.5	20.7	19.9
+ Huge LM	9	21.0	21.3	20.3	20.9	20.1

Table 5: En-De BLEU results. The tuning set is newstest2008–2012. Scores on newstest2014 were computed after the system submission deadline using the released references.

	#iterations	tune	2013	2013 cased	2014	2014 cased
Dense	1	29.1	32.0	30.4	35.6	34.0
Feature-rich	12	37.2	32.5	30.9	36.0	34.5
+ Word class LM	14	35.7	32.3	30.7	–	–

Table 6: Fr-En BLEU results. The tuning set is newstest2008–2012. Scores on newstest2014 were computed after the system submission deadline using the released references.

Word class LM: Unterdessen mehr als 40 Prozent der Bevölkerung *sind* HIV positiv.

In this example, the system without the class-based language model translated the verb twice. In the second translation, the class-based language model prevented this long range disagreement. An analysis of the differences in the translation output of our Fr-En systems showed that the word class language model mainly led to different word choices but does not seem to help grammatically.

4.2 Casing

Our system performed comparatively poorly at casing, as shown in Table 4. In analysis after the evaluation, we found many of these errors related to words with internal capitals, such as “McCaskill”, because the limited recaser we used, which is based on a language model, considered only all lowercase, an initial capital, or all uppercase words. We addressed this issue by allowing any casing seen in the monolingual data. Some words were not seen at all in the monolingual data but, since the target side of the parallel data was included in monolingual data, these words must have come from the source sentence. In such situations, we preserved the word’s original case. Table 7 shows the results with the revised casing model. We gained about 0.24% BLEU for German recasing and 0.15% BLEU for English recasing over our submitted systems. In future work, we plan to compare with a truecased system.

	En-De	Fr-En
Uncased Oracle	20.71	36.05
Conditional Random Field	*19.85*	–
Limited Recaser	19.82	*34.51*
Revised Recaser	20.09	34.66

Table 7: Casing results on newstest2014 performed after the evaluation. The oracle scores are uncased BLEU (%) while all other scores are cased. Submitted systems are shown in *italic*.

5 Negative Results

We experimented with several additions that did not make it into the final submissions.

5.1 Preordering

One of the key challenges when translating from English to German is the long-range reordering of verbs. For this reason, we implemented a dependency tree based reordering system (Lerner and Petrov, 2013). We parsed all source side sentences using the Stanford Dependency Parser (De Marneffe et al., 2006) and trained the preordering system on the entire bitext. Then we preordered the source side of the bitext and the tuning and development data sets using our preordering system, realigned the bitext and tuned a machine translation system using the preordered data. While preordering improved verb reordering in many cases, many other parts of the sentences were often also reordered which led to an overall decrease in translation quality. Therefore, we concluded that this system will require further development before it is useful within our translation system.

5.2 Minimum Bayes Risk Decoding

We further attempted to improve our output by reordering the best 1000 translations for each sentence using Minimum Bayes Risk decoding (Kumar and Byrne, 2004) with BLEU as the distance measure. This in effect increases the score of candidates that are “closer” to the other likely translations, where “closeness” is measured by the BLEU score for the candidate when the other translations are used as the reference. Choosing the best translation following this reordering improved overall performance when tuned on the first half of the newstest2013 test set by only 0.03 BLEU points for the English-German system and 0.005 BLEU points for the French-English system, so we abandoned this approach.

6 Conclusion

We submitted three systems: one constrained Fr-En system, one constrained En-De system, and one unconstrained En-De system. Among all constrained systems, ours performed the best according to uncased BLEU. The key differentiating components of our systems are class-based features, word class language models, and a huge web-scale language model. In ongoing work, we are investigating preordering for En-De translation as well as improved recasing.

Acknowledgements

We thank Michael Kayser and Thang Luong for help with experiments. This work was supported by the Defense Advanced Research Projects Agency (DARPA) Broad Operational Language Translation (BOLT) program through IBM. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575. The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of DARPA or the US government.

References

Anne Abeillé, Lionel Clément, and Alexandra Kinyon, 2003. *Building a treebank for French*, chapter 10. Kluwer.

Timothy C. Bell, John G. Cleary, and Ian H. Witten. 1990. *Text compression*. Prentice-Hall.

Ergun Biçici and Deniz Yuret. 2011. Instance selection for machine translation using feature decay algorithms. In *WMT*.

Christian Buck, Kenneth Heafield, and Bas van Ooyen. 2014. N-gram counts and language models from the common crawl. In *LREC*.

Stanley Chen and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Harvard University, August.

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. 2006. Generating typed dependency parses from phrase structure parses. In *LREC*.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *EMNLP*.

Spence Green, Daniel Cer, Kevin Reschke, Rob Voigt, John Bauer, Sida Wang, et al. 2013a. Feature-rich phrase-based translation: Stanford University's submission to the WMT 2013 translation task. In *WMT*.

Spence Green, Marie-Catherine de Marneffe, and Christopher D. Manning. 2013b. Parsing models for identifying multiword expressions. *Computational Linguistics*, 39(1):195–227.

Spence Green, Sida Wang, Daniel Cer, and Christopher D. Manning. 2013c. Fast and adaptive online training of feature-rich translation models. In *ACL*.

Spence Green, Daniel Cer, and Christopher D. Manning. 2014a. An empirical comparison of features and tuning for phrase-based machine translation. In *WMT*.

Spence Green, Daniel Cer, and Christopher D. Manning. 2014b. Phrasal: A toolkit for new directions in statistical machine translation. In *WMT*.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *ACL*.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *WMT*.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *ICASSP*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL, Demonstration Session*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*.

Shankar Kumar and William Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *HLT-NAACL*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*.

Uri Lerner and Slav Petrov. 2013. Source-side classifier preordering for machine translation. In *EMNLP*.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *NAACL*.

Lucian Vlad Lita, Abe Ittycheriah, Salim Roukos, and Nanda Kambhatla. 2003. tRuEcasIng. In *ACL*.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313–330.

Preslav Nakov, Francisco Guzman, and Stephan Vogel. 2012. Optimizing for sentence-level BLEU+1 yields short translations. In *COLING*.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English gigaword fifth edition, june. *Linguistic Data Consortium, LDC2011T07*.

Jason Smith, Hervé Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the common crawl. In *ACL*. Association for Computational Linguistics, August.

Andreas Stolcke. 2002. SRILM—an extensible language modeling toolkit. In *ICLSP*.

Jakob Uszkoreit and Thorsten Brants. 2008. Distributed word clustering for large scale class-based language modeling in machine translation. In *ACL*.

Ed W. D. Whittaker and Philip C. Woodland. 2001. Efficient class-based language modelling for very large vocabularies. In *ICASSP*.