Alexander Forsyth

6.806 Final Project Writeup

Understanding Shakespeare: Using Statistical Machine Translation To Translate The Plays Of Shakespeare Into Modern English

Abstract

The goal of this project was to translate the plays of Shakespeare into more modern English using statistical machine translation as well as to provide a new 520,000 word parallel corpus for future research. One potential application of this is to use the output translation lattice to obtain a modern English translation in meter (e.g. iambic pentameter). First, a parallel corpus was created by scraping the No Fear Shakespeare website. The Berkeley aligner was used to generate word alignments; KenLM was used to generate the target language (modern English) language model (one using just the target text of the parallel corpus; one also using the English text from the Europarl corpus). Finally, phrase based statistical machine translation was performed using Phrasal. From the two language models used, BLEU scores of 21.769 and 21.598 were obtained respectively on the test data. Human evaluation of the test translations was also conducted. While sample size was small, subjects were generally able to distinguish machine from human translations, especially for longer or multi-sentence lines. The results of this project so far suggest that further work is needed to achieve accurate translations. One limitation of the generated corpus is that some sentence alignments are not one to one, but can be multi-sentence alignments, which makes statistical machine translation more difficult. One area for further work could be to further align the current corpus into individual sentence alignments. After more accurate translations are achieved, future work such as translating into meter could also be explored.

Introduction

This project focused on first providing a new, 520,000 word parallel corpus of Shakespeare's plays in both their original text and in human translated modern English. This project also sought to use this parallel corpus to produce machine translations of Shakespeare for two purposes. First, it sought to produce more accurate, word for word translations of the original text in modern English as opposed to the human translation

that occasionally included liberal translations. And second, it sought to produce an output translation lattice that could later be used to translate into meter (e.g. iambic pentameter). Human translation into meter remains a difficult task and few widely used translations of popular works in meter exist. However, recent work has shown that phrase based machine translation could be used in finite state transducer cascades in order to translate Italian poems into English in meter (Greene et al., 2010). While humans almost uniformly maintain advantages over machines in generating creative language and in translation, machines have some advantages that could aid in translation into meter. E.g., machines can much more easily come up with all possible synonyms of a word and can choose which sequence of potential words fits the desired meter. Therefore, machine translation of Shakespeare could potentially be used to generate a metered, modern English version.

Phrase-based statistical machine translation (Koehn et al., 2003) has been consistently used over the past decade as a common paradigm for machine translation research including the popular open source toolkit, Moses (Koehn et al., 2006). For this project, Phrasal (Cer et al., 2010) was chosen as a similar alternative to Moses for phrase-based statistical machine translation (SMT). It was chosen for its ease of use compared to Moses as well as its favorable performance against Moses in benchmark tests, especially with its default feature set (Cer et al., 2010).

Methods

At a high level, the methods for this project included web scraping the parallel corpus, preparing the corpus for translation, word aligning it, generating the language model, and then performing the phrase-base SMT. Code for the project can be found in the project GitHub (URL shared via Piazza). The only parts of the project not in that repository include the language model files and the Europarl corpus (because of size constraints).

Web Scraping

As mentioned the parallel corpus was scraped from the No Fear Shakespeare website (http://nfs.sparknotes.com/). The website featured the original text of Shakespeare as well as a human translation into modern English. The scraped corpus included all plays of Shakespeare available on the website; it did not include any of the

sonnets. Text was aligned on a "character's line" level, meaning that one character's utterances in the original text could be matched to his translated utterances in the modern text. Unlike most parallel corpora for machine translation, this corpus is not always one to one sentence aligned. For example, when a character's line includes more than one sentence, or when one original sentence is translated to more than one modern sentence, there can be a multi-sentence alignment.

After scraping, the parallel corpus was manually validated by examining randomly selected sentences. Before use, the corpus was cleaned, lowercased and tokenized using the preparation scripts distributed with Phrasal. Finally, the corpus was divided into training, development and test sets with 14 plays in the training set, 4 in the development and 2 in the test.

Machine Translation

Word alignments were performed with the Berkeley Aligner (Liang et al., 2003; DeNero and Klein, 2007) using an algorithm described by the authors as cross-EM training, which jointly trains two conditional Hidden Markov word models. A language model for the target language (i.e. modern English) was created with KenLM (Heafield et al., 2013). The language model was an interpolated n-gram model (Chen and Goodman, 1998) using a maximum of 4-grams with Kneser-Ney smoothing to help with low perplexity (Kneser and Ney, 1995). Two different language models were generated. For the first round of translations, only the target text of the parallel corpus was used to train the language model. For the second round of translations, the target text of the parallel corpus as well as the English portion of the monolingual Europarl corpus was used (Koehn, 2005). Using the corpus, alignments and language model, the machine translation model was trained, tuned, and the test set was decoded all using Phrasal.

Results

Corpus

The parallel corpus generated for this project had approximately 520,000 words covering over half of Shakespeare's plays. This corpus is on the project GitHub and will be released for free for further use. There are a few noteworthy concerns to mention with the current corpus. Speaking character names are included with the sentences; this was known prior to use, but it created a couple problems later on. First, the machine

translation actually translated some character names. E.g., one character, “Sir Toby Belch” was translated into “Sir Toby Vomit”. Also, this formatting confused some human test subjects who thought the names were a part of the sentences they were evaluating. This problem could potentially be solved by compiling a list of character names and by removing these names from the beginnings of all sentences.

The second major issue with the parallel corpus is that sentences are not one to one aligned as mentioned earlier. This could be solved later on by, for example, using a bilingual sentence aligner. There were some more minor issues with the corpus to note. E.g., rarely, words would lose the spacing between them and become concatenated. Solving all of these problems would be necessary before the corpus could be used further.

Translation Data

Original	**Human Translation**	**Machine Translation**
Petruchio: come , where be these gallants ? who 's at home ?	**Petruchio:** whoa ! where is everybody ?	**Petruchio:** come , where are these gentlemen ? who 's at home ?
Malvolio: were not you even now with the countess olivia ?	**Malvolio:** excuse me , weren 't you with countess olivia just now ?	**Malvolio:** weren 't you just now with the countess olivia ?
Sir Toby Belch: thy reason , dear venom , give thy reason .	**Sir Toby Belch:** why are you leaving , my angry friend ?	**Sir Toby Vomit:** your reason , dear poisonous , give your reasonable .
Orsino: give me some music . (music plays) now , good morrow , friends.-- now , good cesario , but that piece of song , that old and antique song we heard last night . methought it did relieve my passion much , more than light airs and recollected terms of these most brisk and giddy-paced times : come , but one verse .	**Orsino:** play me some music . (music plays) good morning , my friends.--have them sing me that song again , cesario , that old-fashioned song someone sang last night . it made me feel better and took my mind off my troubles much better than the silly songs they sing nowadays . please , have them sing just one verse .	**Orsino:** give me some music . (music plays) now , good morning , friends.-- now , good cesario , but that piece of song , that old and ancient song we heard last night . i thought it did rights my feelings much , more than light breezes and recollected terms of these most fresh and giddy-paced times : come , but one verse .

Figure 1 (above). Example translations featuring the original sentences, human translation and machine translation from the translation using the Europarl language model.

Translations were evaluated using the BLEU score. The model without Europarl scored 21.769; the model with Europarl scores 21.598. In addition, some human evaluation was conducted. In the first evaluation with results shown in Figure 2, subjects were given an original Shakespeare line as well as three translations of that line (human, machine without Europarl, and machine with Europarl). Subjects were asked to rate the quality of each of these as "Terrible", "Below Average", "Above Average", or "Great". These scores were converted respectively to numerical scores of 1, 2, 3, and 4. In this evaluation, there were 11 human subjects who each evaluated four different sentences.

Human translation mean	3.341
Machine translation mean	2.182
Machine with Europarl mean	2.318
All translation mean	2.614
All translation standard deviation	0.9050

Figure 2 (above). Average scores for the first human evaluations as well as standard deviation.

In the second round of human evaluation, 6 test subjects were each asked to examine up to 50 short sentences (not all subjects examined all 50 sentences). These 50 sentences were randomly chosen from all sentences in the corpus of length less than 80 characters. Subjects were presented with an original Shakespeare line as well as one translation. The translations were half human and half machine with Europarl translations presented in a random order. Subjects were asked to mark whether they thought a translation was done by a human or by machine (without being asked their preferences for either human or machine translations).

	Count	**Count:Total**
MT marked as HT	72	0.5255
MT marked as MT	65	0.4745
HT marked as HT	106	0.7910
HT marked as MT	28	0.2090

Figure 3 (above) counts of human responses, for example the number of actual machine translations that subjects marked as human translations. Count:total gives the ratio of, for example of the count of machine translations marked as human translations to the total number of machine translations evaluated.

Conclusions

While few conclusions can be drawn from BLEU scores generally, one interesting result was that the score went down after incorporating Europarl into the language model. Assuming BLEU is a measure of "translation quality", this result would be counterintuitive. That is not, however, what BLEU measures. One plausible explanation for this result is that the test corpus translation was similarly structured, perhaps even having been translated by the same person, as the training and development corpora. Thus, BLEU score was maximized when a language model was trained only on the parallel corpus. Training also on Europarl might have many n-grams infrequently used by the human translator.

Definitive conclusions from the human evaluations also cannot be made because the sample sizes were small due to the scope of this project. However, some interesting results can be analyzed. In the first round of evaluations, human translations scored about one entire "rating" above the best machine translations. This result suggests, as expected, that test subjects perceived human translations to still be superior. No definitive conclusion can be made comparing the baseline machine translation to the model incorporating Europarl as their mean ratings differ by less than a standard deviation.

The second round of human evaluation yielded two interesting results. First, subjects essentially flipped a coin when marking machine translations. This suggests that the translations were not often obviously distinguishable. Interestingly, there was not a single sentence that every subject marked as machine translated. This could indicate general similarities in the machine and human translations, subject ignorance of potential differences, or a combination of both.

Overall, these results did not provide conclusive evidence that the machine translations performed better incorporating Europarl into the language model, or that the machine translations could compare favorably to human translations. Further work would be necessary before these translations could, for example, be used as an alternative to human translations or before they could be used to translate into meter. The primary

positive conclusion from this project, and a major focus of it, is that there is now a 520,000 word parallel corpus of Shakespeare's plays freely available for future use.

Future Work

The two major routes for direct improvement of this translation model have been briefly mentioned already. First, a one to one sentence alignment using a bilingual sentence aligner could be generated. Alternatively, long lines could be thrown out of the corpus during cleaning as is done in the Phrasal user guide (Cer et al., 2010). This corpus would then be more in line with traditional machine translation corpora that are sentence aligned. The second route is to improve the corpus by fixing some of the current errors in it. Although they are of slightly different style, Shakespeare's translated sonnets could also be added to the corpus from the No Fear Shakespeare website. Finally, having multiple human translations would likely improve the quality of machine translations, although this would require significant manual investment.

Once machine translations of sufficient quality are available, the most useful application of it would ultimately be to translate into iambic pentameter or another suitable meter. This task is very difficult for humans, but has shown promise with machine translation in prior work (Greene et al., 2010). Finally, the code for this project as well as the 520,000 word parallel corpus will be released for use in any other future work.

Acknowledgements

Foremost, I would like to thank Professor Barzilay for meeting with me throughout the project and for providing very helpful comments and suggestions. I would also like to thank Professor Jaakkola and the TA's, Franck, Karthik, and Tianheng for helping me learn throughout this course.

References

Daniel Cer, Michel Galley, Daniel Jurafsky, and Christopher D. Manning. 2010. Phrasal: a toolkit for statistical machine translation with facilities for extraction and incorporation of arbitrary model features. *Naacl-2010*(June):9–12.

Stanley F. Chen and Joshua T. Goodman. 1998. An Empirical Study of Smoothing Techniques for Language Modeling. (TR-10-98):9.

John DeNero and Dan Klein. 2007. Tailoring Word Alignments to Syntactic Machine Translation. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*:17–24.

Erica Greene, Lancaster Ave, Kevin Knight, and Marina Rey. 2010. Automatic Analysis of Rhythmic Poetry with Applications to Generation and Translation. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP'10)*(October):524–533.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H Clark, and Philipp Koehn. 2013. Scalable Modified Kneser-Ney Language Model Estimation. *Acl*:690–696.

R. Kneser and H. Ney. 1995. Improved backing-off for M-gram language modeling. *1995 International Conference on Acoustics, Speech, and Signal Processing*, 1:181–184.

Philipp Koehn. 2005. Europarl : A Parallel Corpus for Statistical Machine Translation. *MT Summit*, 11:79–86.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation.

Philipp Koehn, Wade Shen, Marcello Federico, Nicola Bertoldi, Chris Callison-Burch, Brooke Cowan, Chris Dyer, Hieu Hoang, Ondrej Bojar, Richard Zens, Alexandra Constantin, Evan Herbst, and Christine Moran. 2006. Open Source Toolkit for Statistical Machine Translation. *Proceedings of ACL*(June):177–180.

Percy Liang, Ben Taskar, and Dan Klein. 2003. Alignment by Agreement.