Towards Automated Translation of Poetry

Michael A.

Introduction

Translating poetry is hard. Without an intimate knowledge of the source and target language, a translation of any text will, without a doubt, be lacking. Translating poetry is even harder: and to do it computationally, almost impossible. This research project explores how we may improve current computational approaches to allow for better-quality translations of poetry.

Why would we want to translate poetry? A couple of reasons come to mind, which specifically motivated my interest in this problem:

- Foreign-language poems translated to English have historically built up the rich tapestry of English literature: where would we be without Dante, Virgil, Homer?
- Machine translation could be used to translate poems to and from languages which have small, if any, translation communities.
- A faithful translation requires the translator to have intimate familiarity with both the source and target languages: a computer could do one of these jobs, allowing for a lighter cognitive load on the translator and perhaps a better-quality translation.
- Translated poems are often worth reading for their own merits, and it would be an interesting creative exercise to analyze machine translated poems.
- The hope that a more computational, quantitative treatment of poetry and poetic analysis will encourage the more scientifically-minded to explore poetry and the beauty contained therein.

Literary theories

Vladimir Nabokov, in the notes of his controversial translation of Aleksandr Pushkin's work 'Eugene Onegin',[1] argues that all translations of poetry will inevitably fall under three categories:

Lexical Translating the basic meaning of words and their order.

Paraphrastic A free version of the original: words and phrases are toyed with — added, removed, or changed — in order to conform to some form that the translator wishes.

Literal Translating as closely as possible the original, with the exact contextual meaning being preserved.

Computationally speaking, the easiest to implement is lexical translation. At most, it requires a good dictionary between the two chosen languages, along with a possible word realignment function. It will not preserve rhyme schemes, vocabulary choices, or poetic meter: but it may provide a useful gloss of the poem for someone only vaguely familiar with the original language.

A literal approach, on the other hand, is currently impossible. It would require a 'true' machine translation system: one which is fully able to understand the choices the author made, and is able to translate these choices into the target language. At this time, a literary translation of any poem is an AI-complete problem, and thus unsolvable until a strong AI system is built — that is, a machine which is as emotionally intelligent as a person.

However, a paraphrastic translation is entirely within the scope of existing machine translation systems. My research has concerned developing hypotheses which would allow current (phrase-based) translation algorithms to make a reasonable attempt at a paraphrastic translation.

Hypotheses

I have developed three hypotheses regarding how current phrase-based statistical approaches could be improved to better translate poetry:

1. Training the language and alignment models on a poetic corpus improves poetic qualities of output translation.
2. Altering the sentence alignment model for poetic works will improve line-by-line translation quality.
3. Poetic characteristics can be preserved by constraints on the hypothesis space, or recovered by post-processing of the output translation.

Statistical methods

It is hypothesized that humans have an internal grammar model, or a 'universal grammar',[2] which allows them to learn language in a way that other animals cannot. But exactly what this grammar is eludes researchers: so, try as we might, we cannot computationally emulate it. Instead, we can approximate human use of language using statistical methods.

Say we want to translate a foreign sentence f into an English sentence e. Then by a probabilistic model, we want to maximise the probability that a given English sentence is the correct translation of the foreign sentence: that is, maximise $\Pr(e|f)$.[4] Modelling this probability distribution is difficult, so we simplify the problem by applying Bayes' theorem:[3]

$$\Pr(e|f) = \Pr(e)\Pr(f|e)$$

This gives us two variables to maximize: $\Pr(e)$, the *language model*, and $\Pr(f|e)$, the *alignment model*.

Alignment model

An alignment model is used to translate words or phrases from a foreign language into our target language. With a phrase-based translation system, we segment our input and output text into aligned phrases, and use these to build a probability distribution. We estimate this probability distribution by expanding $\Pr(f|e)$:

$$\Pr(\bar{f}_1^I|\bar{e}_1^I) = \prod_{i=1}^{j} \phi(\bar{f}_i|\bar{e}_i)\, d(\text{start}_i - \text{end}_{i-1} - 1)$$

I hypothesize that we can separate poems into two different classes, which should have different alignment models applied to them, specifically with reference to sentence boundaries:

1. Poems whose sentences can be treated in the same way as prose: standard sentence boundaries can be used to segment sentences.
2. Poems whose sentence boundaries are marked by line-endings, rather than punctuation.

Language model

In machine translation, a language model is needed to make the output sound fluent in the target language. To estimate its probability distribution, we take a monolingual corpus and use n-grams to estimate the probability of word sequences.

Language models are very sensitive to their training corpus: the training corpus should be in the same domain and on the same topic as the items to be translated. Therefore, I hypothesize that training the language model on a poetic corpus will increase the poetic qualities of a translation. Part of my research involved building a poetic corpus from the Project Gutenberg collection, which is large enough to feasibly use to train real models.

Preserving poetic qualities

The previous two hypotheses may improve poetic qualities of translations, but they may not necessarily preserve the poetic techniques that were used in the original poem. There are two possible ways to approach this:

1. Hypothesis constraints, which can preserve such features as line length (syllable and word), rhyme, or meter.[5]
2. Post-processing: in effect, building a translation system from non-poetic to poetic English.

However, we are still faced with the problem of how to identify poetic techniques computationally. This is more difficult: nonetheless, some methods have already been developed to, for example, discover rhyme schemes.

Conclusion

Each of the hypotheses I have suggested for improving phrase-based statistical models of machine translation are good candidates for future work, as they would all be relatively easy to integrate into existing machine translation systems such as Moses.

Translating poetry is hard, there is no doubt about that. It is my hope that it will one day improve to the point that we may be able to alter Robert Frost's famous quotation:

Poetry is that which is lost in translation,
Unless we use a computational calculation.

References

[1] V. Nabokov. *Eugene Onegin: A Novel in Verse by Aleksandr Pushkin. Translated from the Russian*. 1964.

[2] N. Chomsky. *Aspects of the Theory of Syntax*. 1965.

[3] T. Bayes & R. Price. *An Essay towards solving a Problem in the Doctrine of Chance*. 1763.

[4] P. Koehn. *Statistical Machine Translation*. 2010.

[5] D. Genzel & J. Uszkoreit & F. Och. *"Poetic" Statistical Machine Translation: Rhyme and Meter*. 2010.