Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts
AAAI Technical Report CR-13-01

Crowdsourcing Translation by Leveraging Tournament Selection and Lattice-Based String Alignment

Julien Bourdaillet⋆ and **Shourya Roy**† and **Gueyoung Jung**⋆ and **Yu-An Sun**⋆
Xerox Innovation Group
⋆ 800 Philips Road, Webster, NY; † Xerox Research Centre India, Bangalore, India
{ firstname.lastname }@xerox.com

Abstract

Crowdsourcing translation tasks typically face issues due to poor quality and spam translations. We propose a novel method for generating large multilingual text corpora leveraging Tournament Selection and Lattice-Based String Alignment without requiring expert involvement or *Gold data*. We use crowdsourcing for gathering a set of candidate translations of a given source sentence. A crowd sourced Tournament Selection allows to prune this set and keep an n-best list. Finally, a lattice-based string alignment allows to compose parts of these n-best translations to generate a consensus translation. As a work-in-progress, the evaluation is still ongoing, but we are confident in the potential for this method to generate translations of good quality.

Introduction

We propose a method for translating sentences from one language to another using crowdsourcing and addressing the above issues. It relies on crowdsourcing the translation process, data cleaning via tournament selection (i.e., removing poor translations), and generating the consensus translation of each sentence using a lattice-based word alignment. (Sun, Roy, and Little 2011) proposed a tournament selection based quality control process for crowdsourced translations. This technique allows a minority of good quality translations to prevail over the majority that are incorrect in some way. Though this work showed that a good translation eventually emerges as the winner, the technique had to perform several steps of tournament selection to identify the best winner, making the overall process less efficient with respect to time and cost. In addition, the eventual winner will always be one of the crowd translations even if it is not an accurate translation. We extend that work by bringing in the notion of lattice-based string alignment. This approach has been used in Machine Translation to generate a translation of better quality from several machine generated translations (Bangalore, Bordel, and Riccardi 2001).

In our approach, we embed a sentence to translate in a microtask that is translated several times by different crowdworkers (see Figure 1). Then a crowdsourced pairwise Tournament Selection is organized to select the best sentences.

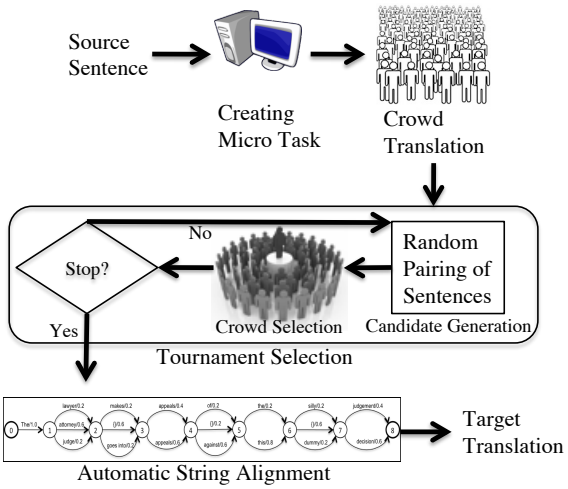

Figure 1: Block diagram of our approach

Iterative execution of this step takes out poor quality translations (primarily by the novice, incompetent or spamming workers) and thereby leaving only potentially good translations. The chosen translations are aligned using a word-based edit distance to form a string lattice which is searched in order to generate the best target translation by composing parts of candidate translations. In the end, a target consensus translation of the source sentence is obtained.

Our Approach

Crowd Translation

For a given translation task, crowdsourcing enables access to a huge number of crowd workers who have different levels of translation skill. The quality of the translation results vary a lot - in particular, when the sentence is difficult to translate. It can be evaluated from "very well" as a professional to "very bad" or even "not a translation". To overcome this problem, our system asks multiple crowd workers to translate the same single sentence. We obtain a set of translations in the target language for a given source sentence. Although it is possible that none of the translations as a whole is correct, the larger the translation set is, the more likely the set contains well-translated parts of the source sentence. Hence,

it is possible to combine such well-translated parts to form a whole target sentence that is a high quality translation of the source sentence.

Tournament Selection Using the Crowd

Given a set of translations from the crowd workers, this step selects the n best translations from this set. Automatically selecting such best translations is not trivial. In Statistical Machine Translation, this is achieved by associating probabilities that reflect the quality of each candidate, while generating them. In our case, we do not have such probabilities associated with each crowdsourced translation. To bypass that, we use a crowdsourced tournament selection approach (Sun, Roy, and Little 2011). The idea of tournament selection is to select the best candidate among a given population by organizing a multi-round tournament.

For this, we create random pairs of candidate translations and ask crowdworkers to decide which one is a better translation of the source sentence. This is repeated for n crowdworkers resulting in a set of n candidate translations which are a subset of the initial crowd translations. (Sun, Roy, and Little 2011) repeated this step m times to show that only good translations survive these steps, and that the best translation becomes the majority at the end. This iterative process can be stopped when the population has been sufficiently pruned (for example to 20-30% of its original size) while remaining sufficiently diverse. The "Stopping Condition" is a configurable parameter and depends on many factors such as the expertise of the crowd, the difficulty of the translation task, the number of initial/candidate translations etc. In our experiments, we found that 2 steps of Tournament Selection are sufficient to obtain good results.

Automatic String Alignment

Given a set of n top ranked candidate translations, this step composes parts of these translations to form a clean and proper translation of the source sentence. Our approach automates this step using a lattice-based string alignment. This technique was originally proposed in (Bangalore, Bordel, and Riccardi 2001) to combine translations resulted from multiple machine translation systems. In our case, the set of candidate translations results from crowdworkers and is noisier than machine translation systems, however, it potentially contains better translations since human translated them. Nevertheless, the crowdsourced Tournament Selection has permitted to clean the set which is supposed to contain only reasonably clean candidate translations.

To align the candidate translations, we rely on the method of progressive multi string alignment proposed initially in bioinformatics (Feng and Doolittle 1987). We designed a word-based edit distance that allows to compare candidate translations using insertions, deletions and substitutions of words, and that aligns preferentially words that share an identical Part-Of-Speech. It is then straightforward to compute a distance matrix among all candidate translations. An iterative process aligns the closest candidates in pair-wise manner according to the distance matrix. At the end, we obtain a multi-alignment between all candidate sentences (see Figure 2).

The	lawyer	makes	appeal	of	this	silly	judgment
The	attorney		appeals		the		decision
The	attorney		appeals	against	this	dummy	decision
The	attorney		appeals	against	this		decision
The	judge	goes into	appeal	against	this		judgment

Figure 2: Multi alignment of 5 translation candidates resulted from the crowdsourced Tournament Selection

From this alignment, a lattice is built by merging the edges that share a common word (see Figure 3). Each edge is associated with a probability by computing the ratio of candidate sentences that share the word labeling it. Finally, a shortest path search is used to compute the best translation possible from the lattice. At the end, we obtain a translation of the source sentence that is composed of parts from several different candidate sentences. This approach leverages the candidate translations from the crowdworkers in order to generate a new target translation that was never proposed by them.

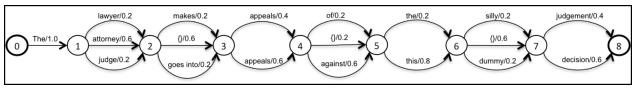

Figure 3: Lattice built after merging the identical edges, and computing a probability for each edge. It will permit to generate the final translation by searching the best path

References

Bangalore, S.; Bordel, G.; and Riccardi, G. 2001. Computing consensus translation from multiple machine translation systems. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding. ASRU '01*, 351 – 354.

Feng, D.-F., and Doolittle, R. 1987. Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. *Journal of Molecular Evolution* 25:351–360.

Sun, Y.-A.; Roy, S.; and Little, G. D. 2011. Beyond independent agreement: A tournament selection approach for quality assurance of human computation tasks. In *Proceedings of the AAAI Workshop on Human Computation.*