Are Two Heads Better than One? Crowdsourced Translation via a Two-Step Collaboration of Non-Professional Translators and Editors

Rui Yan, Mingkun Gao, Ellie Pavlick and Chris Callison-Burch

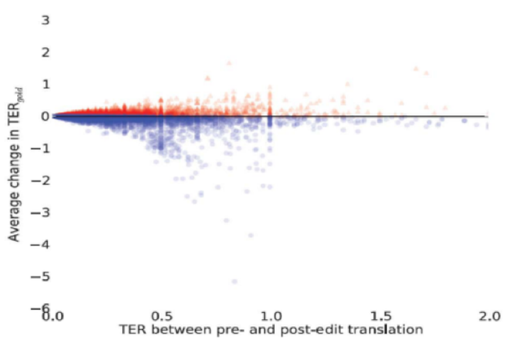
Computer and Information Science Department,
University of Pennsylvania, Philadelphia, PA19104, USA
{ruiyan, gmingkun, epavlick, ccb} @seas.upenn.edu

Crowdsourcing Preliminary

- **A viable mechanism for creating large-scale training data for Natural Language Processing techniques, i.e., machine translation, etc.**
  - Low cost
  - Fast turn-around time
  - Especially useful under the scenario of aiding "low resource" languages.
- **Potential Pitfalls**
  - Non-professionals
  - Low quality
- **Solution**
  - Automated quality control!
  - Two-Step collaboration between translators and post-editors based on graph ranking

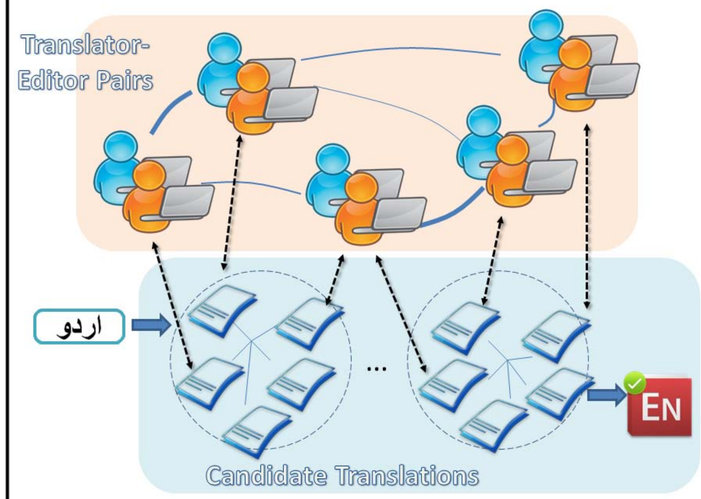
Crowdsourcing Translation

- **Setups**
  - Data set: 1,792 Urdu sentences, paired with English translations.
  - Each Urdu sentence was translated redundantly by 3 distinct translators
  - Each translation was edited by 3 separate native English speakers as post-editors
  - 52 Turkers took part in the translation task, each translating 138 sentences
  - 320 Turkers participated in editing task, average 56 sentences edited each
- **Some editors make improvements but it is tricky to automatically identify the good ones**
  - Agreement picks out lazy editors!



Graph based Crowdsourcing Translation Model

- **Framework**



The 2-step collaborative crowdsourcing translation model based on graph ranking framework includes three sub-networks. The undirected links between users denotes translation-editing collaboration. The undirected links between candidate translations indicate lexical similarity between candidates. A bipartite graph ties candidate and Turker networks together by authorship (to make the figure clearer, some linkage is omitted). A dashed circle indicates the group of candidate translations for a single source sentence to translate.

- **Problem formulation**
  - Given a set of candidate translation for a particular source sentence, the goal is to choose the best output translation.
  - We form two graphs: the first graph (G_T) represents Turkers (translator/ editor pairs) as nodes; the second graph (G_C) represents candidate translated and edited as nodes.
  - The two graphs (G_T and G_C) are combined as sub-graphs of a third graph (G_{TC}). Edges in G_{TC} connect author pairs (nodes in G_T) to the candidate they produced (nodes in G_C).

$$G = (V, E) = (V_T, V_C, E_T, E_C, E_{TC})$$
$$G_C = (V_C, E_C)$$
$$G_T = (V_T, E_T)$$
$$G_{TC} = (V_{TC}, E_{TC})$$
$$V_{TC} = V_T \cup V_C$$

- **Inter-Graph Ranking**
  - A candidate sentence is important if 1) it is similar to many other proposed candidates and 2) it is authored by better qualified translators and/or post-editors
  - A translator/editor pair is better qualified if 1) the editor is collaborating with a good translator and vice versa and 2) the pair has authored important candidate translations

Introducing the saliency scores for candidate sentence and turker pairs, we can formulae as:

- Homogeneity: $\mathbf{c} \propto M^T\mathbf{c}, \quad \mathbf{t} \propto N^T\mathbf{t}$
- Heterogeneity: $\mathbf{c} \propto \hat{W}^T\mathbf{t}, \quad \mathbf{t} \propto \bar{W}^T\mathbf{c}$
- Computing steps: 1) compute the saliency scores of candidates and then normalize and 2) compute the saliency scores of turker pairs and then normalize. Repeat until convergence.

$$\mathbf{c}^{(n)} = (1-\lambda)M^T\mathbf{c}^{(n-1)} + \lambda\hat{W}\mathbf{t}^{(n-1)} \qquad \mathbf{t}^{(n)} = (1-\lambda)N^T\mathbf{t}^{(n-1)} + \lambda\bar{W}\mathbf{c}^{(n-1)}$$
$$\mathbf{c}^{(n)} = \mathbf{c}^{(n)}/||\mathbf{c}^{(n)}||_1 \qquad \mathbf{t}^{(n)} = \mathbf{t}^{(n)}/||\mathbf{t}^{(n)}||_1$$

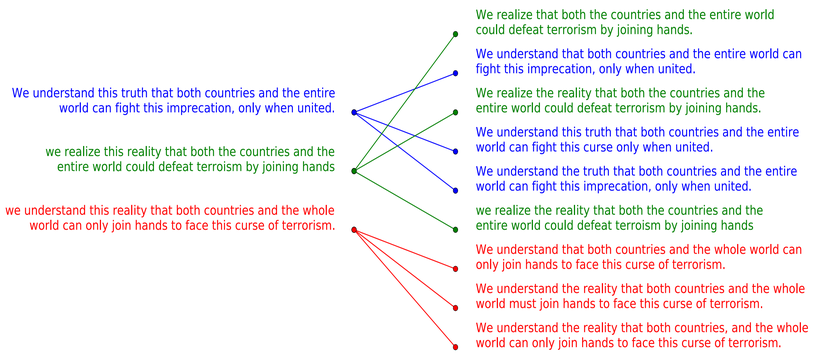
- **Intra-Graph Ranking**
  - Pagerank schema:

$$\mathbf{c} = \mu M^T\mathbf{c} + (1-\mu)\frac{\mathbf{1}}{|V_C|} \qquad \mathbf{t} = \mu N^T\mathbf{t} + (1-\mu)\frac{\mathbf{1}}{|V_T|}$$

Experiment and Evaluation

- **Evaluation metric**
  - BLEU: Bilingual Evaluation Understudy score
  - 4 references from professional translator as ground truth set
- **Baselines**
  - Random
  - Oracle: based on segment level and Turker level separately, and both based on translation only and translation plus post-editings
  - Lowest TER: to select the translation with the minimum average TER
  - Linear combined regression (results directly reported*)
  - Graph based ranking based on translation only
  - Graph based ranking based on translator/editor collaboration
- **Experimental results**

Example Rankings



Conclusions

- **Contributions**
  - An analysis of the difficulties posed by a 2-step collaboration between editors and translators in crowdsourcing environment
  - A graph based algorithm for quality control in selecting the best translation