Translating a Language You Don't Know in the Chinese Room

Ulf Hermjakob, Jonathan May, Michael Pust, Kevin Knight • USC Information Sciences Institute • ulf@isi.edu

Chinese Room Objectives

- Machine translation thirsts for bitext, especially in-domain.
- Low-resources languages are low on bitext, especially in-domain.

Primary Objectives

- Enable people to translate from low-resource language to English, even without any prior knowledge of source language.
- Build in-domain bitext for tuning, ideally some more for training.

Secondary Objective

- Support computational linguists in identifying challenges of a specific low-resource language.

Approach

- Reuse machine translation resources such as t-tables and special modules (e.g. for quantities, named entities) to build a glossing tool to support human translators.
- Allow user to explore alternative translations.
- Combine artificial intelligence and human intelligence.

Some Challenges

- Foreign scripts can present a massive cognitive barrier.
  ياپونىيە فۇكۇشىما 1-يادرو ئېلېكتىر ئىستانسىسىنىڭ تۆت گېنراتورلار گۇرۇپپىسى
  Solution: Universal romanizer *uroman* (Ulf Hermjakob et al., ACL 2018).
  yaponie fukushima 1-yadro elektir istansisining toet genratorlar guruppisi
- Inconsistent spelling for many low-resource languages
  due to dialects, lack of spelling standards, lack of education.
  Solution: Multiple indexing methods to find matching words.

Features of the *Chinese Room*

中文室

- Glosser accommodates a variety of NLP and source language resources.
- User can explore alternative translations.
- Grammar support (such as prefixes, suffixes, function words).
- Optional romanization of source text.
- Robust to spelling variations.
- Optional confidence levels.
- Propagation of user translations.
- Dictionary search function (allowing regular expressions).
- User accounts with login, password, worksets, separate workspaces.
- Web-based.

Experiments

- Built *Chinese Rooms* for Bengali, Hungarian, Kinyarwanda, Oromo, Sinhalese, Somali, Swahili, Tagalog, Tigrinya, Uyghur.
- Trained more than 20 people to use the *Chinese Room*.
- Successfully improved MT system for low-resource languages.

This work has been published as *Translating a Language You Don't Know in the Chinese Room* in the Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). Demo track. Melbourne. July 2018.