Multi Language Support for Virtual Assistants

Arabic/Spanish case study

Sierra Kaplan-Nelson, Max Farr, Mehrad Moradshahi
Motivation

- People expect to be able to talk to their VA the way they talk to a friend (conversational)
- 422 million native Arabic speakers
- 483 million native Spanish speakers
- VAs will be how people interact with the web and find information (right now non English speakers are at a severe disadvantage)
State of the Art - Translated VA

- In previous works, data collection is mainly done using human annotators and translators to verify quality.

- Instead, we use automated translation for our task, which works because semantic parsing is more robust to translation noise.
Problems

- Almond currently works for Farsi, Italian and Chinese in the hotels & restaurants domains
- Extending Almond in other languages to other domains
- Data collection for languages without translated structured data (Yelp is in 15 languages)
- Automate Google translate corrections
- Generate colloquial sentences in other languages
State of the Art - Arabic conversational

Google has Saudi and Egyptian Arabic and is adding new Arabic dialects as of Dec. 2019

Siri only supports Modern Standard Arabic and not dialect

Research ML agent to translate dialects to MSA

Arabic and AI: Why voice-activated tech struggles in the Middle East

- Arabic dialect includes many portmanteaus (words created by combining other words) and often exclude separate prepositions
- Arabic dialects haven’t been written down until recently
State of the Art - Arabic conversational
Plan - Colloquial Arabic

- Generate colloquial sentences in Levantine Arabic
  - Translate Levantine sentences into Modern Standard Arabic using neural network
- Incorporate corpuses of named entities in Arabic into model
- Test with native Arabic speaker to get qualitative data
Plan

- Reproduce existing results
- Extending Almond in other languages to other domains - Arabic & Spanish
- Data collection for languages without translated structured data - test new data collection strategies
- Explore Genie training methods
- Add the mBART framework for translation correction
- Implement translation of colloquial Arabic
What We Expect to Learn

- How easy is it to build a generalizable system that people can add to to improve results for specific languages/dialects across domains
- Compare different data collection strategies/platforms (Wikipedia, social media, etc.) versus structured data
- Compare autoencoder training strategies for Arabic dialects
  - Do our strategies generalize to cross-language translation?
- Evaluate mBART’s capabilities for correcting low-quality Google Translate results
Demo

- Setting up a restaurant/hotel reservation in Spanish and Arabic
- Showing a short colloquial Arabic transaction
- Graphs comparing how well each ML method perform at correcting translations
<table>
<thead>
<tr>
<th>Week</th>
<th>Sierra</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 5</td>
<td>Reproduce existing results</td>
<td>Reproduce existing results</td>
</tr>
<tr>
<td>Week 6</td>
<td>Baseline for Arabic/ Spanish for restaurant domain, baseline on hotels domain</td>
<td>Dataset/ parameter collection for Arabic and Spanish</td>
</tr>
<tr>
<td>Week 7</td>
<td>Build model to translate between dialects and MSA, baseline results</td>
<td>Automating sentence-correction for Arabic and Spanish</td>
</tr>
<tr>
<td>Week 8 (paper deadline)</td>
<td>Improve results from 7, try adding named entities corpus in Arabic</td>
<td>Expanding current work to other domains</td>
</tr>
<tr>
<td>Week 9</td>
<td>Discussion on future work</td>
<td>Discussion on future work</td>
</tr>
<tr>
<td>Week 10</td>
<td>Wrap up the project and celebrate!</td>
<td>Wrap up the project and celebrate!</td>
</tr>
</tbody>
</table>