Project Proposal

CS 6320.002: Natural Language Processing

Team
Siddhartha Sahai - sxs180170
Manas Bundele - mmb180005
Randeep Ahlawat - rsa190000

Part 1:

Code-switching of two more languages is pretty common in a multi-lingual society where the user uses a combination of his mother tongue(host) and other foreign language. This gives rise to a completely new mixed language that neither completely follows the rules of host language nor the foreign language. Instead, mostly, the elements of foreign language get morphologically transformed as per the host language.[2] This poses a new challenge in machine translation as its first necessary to understand which word belongs to which language[2] and then translate them to a standard language(target language), also keeping in mind the grammatical rules for the target language.

Code-switching of Hindi and English where the speakers follow the syntax of Hindi but borrow some words or phrases from English[1] and use Romanized text to depict the entire sentence has become quite common nowadays on social networking sites such as twitter, facebook, etc. Also, when one language is transliterated into another language, the same word can be written into multiple ways, based on the choice or pronunciation of the individual and thus normalization of all the possible words into a single word becomes an overhead to the task of translation. The main bottleneck for this problem is the lack of gold parallel data between code-mixed language and standard language.[1] Training a Neural Machine Translation model from scratch requires huge amount of data as well as resources. Thus, we intend to use transfer learning to translate a low-resource code-mixed language into a target language.

Part 2:

As of August 2018, according to the data of internetworldstats.com, the Internet has four-hundred fifty million
English speaking users out of one billion five hundred million total users. This means that the market for English language is slightly less than one third of the total market. In other terms, most current approaches to information extraction exploiting social media and user-generated content (UGC), that are predominantly developed for English, are working with a mere third of the total data available.[1]

In recent years, the use of code-switched language of Hindi and English has become quite predominant among the Indian youth. This data can be used to leverage more accurate searches in search engine, improved question answering, sentiment analysis, humor or sarcasm detection (which is quite common on social networking sites but is often misinterpreted due to its hybrid nature). Since dealing with hybrid text is a relatively more complex and nuanced task as compared to dealing with monolingual languages, approaches used for such standard languages result in loss of information and lead to misinterpretations. Thus, there is a dire need to develop techniques that can deal with multilingual languages effectively, as

per our knowledge. We believe approaches applied specifically to Hindi - English hybrid sentences will assist in tasks involving other multilingual code-switched languages.

Part 3:

1. Mrinal Dhar, Vaibhav Kumar, Manish Shrivastava, 2018. Enabling Code-Mixed Translation: Parallel Corpus Creation and MT Augmentation Approach. Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing, August 2018, Santa Fe, New Mexico, USA, ACL, pp:131–140.

2. Sinha, R.M.K. and A. Thakur, 2005. Machine translation of bi-lingual Hindi-English (Hinglish) text. Proceeding of the 10th Conference on Machine Translation, Sept. 13-15, MT-Archive, Phuket, Thailand, pp: 149-156.

3. Gustavo Aguilar, Thamar Solorio, From English to Code-Switching: Transfer Learning with Strong Morphological Clues. https://arxiv.org/abs/1909.05158v1

4. [Firat *et al.*, 2016] Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073*, 2016.

5. Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data Augmentation for Low-Resource Neural Machine Translation Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'17)