Multilingual Sentiment Analysis for Twitter

George Cooper and Scott Simpson

Goals

- Create methodology for building sentiment analysis tools for twitter for languages that don't have sentiment-specific resources but do have other resources (bilingual dictionary, machine translation software, POS tagger). Test methodology using Spanish tweets.

Methodology

- Compare two approaches:
  - Use machine translation to translate input documents into English and use an English sentiment analysis system (NRC-Canada)
  - Adapt English resources to build a sentiment analysis system in the target language

Translating Input Documents

- Use Google Translate
- Clean tweets in preparation for translation:
  - Remove repeated letters (e.g. “aaawesoooome” -> “awesome”)
  - Split multi-word hashtags (e.g. “#thebestever” -> “#the best ever”)
  - Correct transposed letters (e.g. “hte” -> “the”)

Adapting English Resources

- Translate training data from English into target language using Google Translate (with cleanup)
- Generate new training data in target language by searching for positive and negative emoticons and hashtags
- Translate sentiment lexicons using a bilingual dictionary
- Translate negation word list using a bilingual dictionary

Completed Tasks

- English Sentiment Analysis tool built
- List of positive and negative Spanish hashtags built
- All resources acquired (except for a few that could be useful for tweet cleanup)
- Translated negation word list

In-Progress Tasks

- Queries currently running for Spanish tweets with positive and negative emoticons and hashtags
- Automatically translating sentiment lexicons from English into Spanish
- Writing code to clean tweets in preparation for machine translation

Remaining Tasks

- Use machine translation to translate training data from English into Spanish
- Run Spanish experiments using translated training data, translated inputs, and training data from tweets with specific emoticons and hashtags.

Preliminary Results

- English twitter sentiment analysis tool achieves 65.12 averaged F-score (compared to 69.02 reported in paper)

Predicted Results

- We predict that there will be a modest drop in accuracy for Spanish sentiment analysis tools
- Don’t know whether translating input tweets into English or adapting English resources will yield better accuracy