SMS Text Normalization using Statistical Machine Translation

How to translate your daughter's text messages

http://nlp.stanford.edu/sms/translate.php

su nlp grp is gr8
stanford university nlp group is great

ovrvw

- Text message normalization is the process of translating "text speak" into plain English.
- A simple dictionary lookup and replace approach is often unsatisfactory.
  E.g. "do u noe how 2", which one would the dictionary pick: to, too, two?
- Successful translation needs understanding of context and domain knowledge.
  E.g. wishing u a happy new yr! -> wishing you a happy new year!
  is dat yr house? -> is that your house?
- We pose this as a statistical machine translation (SMT) task from SMS to English.
- We built a state-of-the-art system using standard machine translation (MT) components.

challngz

- Disambiguation decisions: 2 -> to, too, two; wk -> week, work.
- To normalize or not? sat -> sat (v.), saturday (n.).
- No formal convention for SMS words: "e" for "the" in Singapore, "da" in the US.
- Colloquial terms: e.g. lea, lor, mah in Singapore.
- Region specific disambiguation: e.g. "mrt" could be "market", but in Singapore refers to a public transport system.
- Topic specific disambiguation: e.g. "asl" in a chatting scenario is "age, sex, language", while in another discussion it could be "American sign language".
- Spelling variations: e.g. wanna -> wana, wannna; people -> ppl, pplle.

smt mdl

Our model follows the standard phrase-based SMT formulation to translate from text speak to English:

\[ E = \arg \max_{E \in \text{English}} \Phi(F|E) \cdot d(F,E) \cdot P(E) \cdot \omega(\text{length}(E)) \]

phrase translation probability  distortion model  language model  word cost

resltts

Comparing to Previous Work

Performance on SMSes from different regions

Karthis Raghunathan
rkarthis@cs.stanford.edu

Stefan Krawczyk
stefank@cs.stanford.edu

Christopher Manning
manning@stanford.edu