Issues in Arabic MT

Alex Fraser
USC/ISI

ISI Arabic System for 2003 TIDES Evaluation

- Alignment Template Approach (Och and Others at RWTH Aachen)
- Maximum BLEU training (Och, ACL 2003)
- Customization of Training for Arabic System
  - Model and Search are not Arabic dependent (currently)
- Top Scoring System

Character Encoding and Normalization

- Arabic UTF-8 reduced to CP-1256 character set (8-bit MS-Windows encoding)
  - Handle non-Arabic characters that look similar
  - Numbers
- Normalization is important
  - Strip Kashida, vowels, Shadda
  - Normalize Alef variants, Alef Maqsura/Yeh, Heh/The Marbuta, Hamza variants

Morphology

- Simple morphological segmentation did not improve performance at large training sizes
  - MT Extensions to UMass Light Stemming for IR (Larkey et al., SIGIR 2002)
  - Modified Buckwalter Stemmer (LDC), conservative stems (Xu, Fraser, Weischedel, SIGIR 2002)
  - Space-separated Arabic strings are already translated as consecutive-word phrases with baseline system
- Used Buckwalter Stemmer and Gloss for unknown words

Training on long sentences

- Realignment of sentences of length > 45 tokens on chunk level
- Virtually all data can be used for training (93M words English, 82M words Arabic).
- English chunks are projected to Arabic
- IBM Model 1 Viterbi word alignment is used to project high precision chunk breaks from English to Arabic
- Dynamic programming search for best chunk projection

Error Analysis

- Verbal movement and form
  - VSO ordering
  - Tense
- NP structure
- Missing 'to be' in present tense
  - Also causes spurious ‘to be’
- PRO
- These are all syntactic problems
- Also Important: Named Entities, Unknown Words

Future

- More parallel data – 1 billion words
  - More in-domain data
- More test sets
- Named Entity list
- Research on Syntax