The CMU Arabic-to-English Statistical MT System

Alicia Tribble, Stephan Vogel

Language Technologies Institute
Carnegie Mellon University

The Data

- For translation model:
  - UN corpus: 80 million words UN
  - Ummah
  - Some smaller news corpora
- For LM
  - English side from bilingual corpus: Language model should have seen the words generated by the translation model
  - Additional data from Xinhua news
- General preprocessing and cleaning
  - Separate punctuation mark
  - Remove sentence pairs with large length mismatch
  - Remove sentences which have too many non-words (numbers, special characters)

The System

- Alignment models: IBM1 and HMM, trained in both directions
- Phrase extraction
  - From Viterbi path of HMM alignment
  - Integrated Segmentation and Alignment
- Decoder
  - Essentially left to right over source sentence
  - Build translation lattice with partial translations
  - Find best path, allowing for local reordering
  - Sentence length model
  - Pruning: remove low-scoring hypotheses

Some Results

- Two test sets: DevTest 203 sentences, May2003
- Baseline: monotone decoding
- RO: word reordering
- SL: sentence length model

	DevTest	DevTest	May 2003
	NIST	Bleu4	NIST
Baseline	8.59	0.385	8.95
RO	9.02	0.441	9.26
RO + SL	9.24	0.455	?

Questions

- What's specific to Arabic
- Encoding
- Named Entities
- Syntax and Morphology
- What's needed to get further improvements

What's Specific to Arabic

- Specific to Arabic
  - Right to left not really an issue, as this is only display
    Text in file is left to right
  - Problem in UN corpus: numbers (Latin characters) sometimes in the wrong direction, eg. 1997 -> 7991
- Data not in vocalized form
  - Vocalization not really studied
  - Ambiguity can be handled by statistical systems

Encoding and Vocalization

- Encoding
  - Different encodings: Unicode, UTF-8, CP-1256, romanized forms not too bad, definitely not as bad as Hindi;-)
  - Needed to convert, e.g. training and testing data in different encodings
  - Not all conversion are loss-less
  - Used romanized form for processing
- Converted all data using 'Darwish' transliteration
  - Several characters (ya, allef, hamzda) are collapsed into two classes
  - Conversion not completely reversible
- Effect of Normalization
  - Reduction in vocabulary: ~5%
  - Reduction of singletons: >10%
  - Reduction of 3-gram perplexity: ~5%

Named Entities

- NEs resulted in small but significant improvement in translation quality in the Chinese-English system
- In Chinese: unknown words are splitted into single characters which are then translated as individual words
- In Arabic no segmentation issues -> damage less severe
- NEs not used so far for Arabic, but started to work on it

Language-Specific Issues for Arabic MT

- Syntactic issues: Error analysis revealed two common syntactic errors
  - Verb-Noun reordering
  - Subject-Verb reordering
- Morphology issues: Problems specific to AR morphology
  - Based on Darwish transliteration
  - Based on Buckwalter transliteration
  - Poor Man's morphology

Syntax Issues: Adjective-Noun reordering

- Adjectives and nouns are frequently reordered between Arabic and English
- Example: EN: 'big green chair'
  AR: 'chair green big'
- Experiment: identify noun-adjective sequences in AR and reorder them in preprocessing step
  - Problem: Often long sequences, e.g. N N Adj Adj N Adj N N
  - Result: no improvement

Syntax Issues: Subject-Noun reordering

- AR: main verb at the beginning of the sentence followed by its subject
- EN: order prefers to have the subject precede the verb
- Example: EN: 'the President visited Egypt'
  AR: 'Visited Egypt the President'
- Experiment: identify verbs at the beginning of the AR sentence and move them to a position following the first noun
  - No full parsing
  - Done as preprocessing on the Arabic side
  - Result: no effect

Morphology Issues

- Structural mismatch between English and Arabic
  - Arabic has richer morphology
  - Types Ar-En: ~2.2 : 1
  - Tokens Ar-En: ~ 0.9 : 1

Tried two different tools for morphological analysis:

- Buckwalter analyzer
  - http://www.xrce.xerox.com/competencies/content-analysis/arabic/info/buckwalter-about.html
  - 1-1 Transliteration scheme for Arabic characters
- Darwish analyzer
  - www.cs.umd.edu/Library/TRs/CS-TR-4326/CS-TR-4326.pdf
  - Several characters (ya, alef, hamza) are collapsed into two classes with one character representative each

Morphology with Darwish Transliteration

- Addressed the compositional part of AR morphology since this contributes to the structural mismatch between AR and EN
- Goal was to get better word-level alignment
- Toolkit comes with a stemmer
- Created modified version for separating instead of removing affixes
- Experiment 1: Trained on stemmed data
  - Arabic types reduced by ~60%, nearly matching number of English types
  - But loosing discriminative power
- Experiment 2: Trained on affix-separated data
  - Number of tokens increased
  - Mismatch in tokens much larger
- Result: Doing morphology monolingually can even increase structural mismatch

Morphology with Buckwalter Transliteration

- Focused on DET and CONJ prefixes:
  - AR: 'the', 'and' frequently attached to nouns and adjectives
  - EN: always separate
- Different spitting strategies:
  - Loosest: Use all prefixes and split even if remaining word is not a stem
  - More conservative: Use only prefixes classified as DET or CONJ
  - Most conservative: Full analysis, split only can be analyzed as a DET or CONJ prefix plus legitimate stem
- Experiments: train on each kind of split data
- Result: All set-ups gave lower scores

Poor Man's Morphology

- List of pre- and suffixes compiled by native speaker
- Only for unknown words
  - Remove more and more pre- and suffixes
  - Stop when stripped word is in trained lexicon
- Typically: 1/2 to 2/3 of the unknown words can be mapped to known words
- Translation not always correct, therefore overall improvement limited
- Result: this has so far been (for us) the only morphological processing which gave a small improvement

Experience with Morphology and Syntax

- Initial experiments with full morphological analysis did not give an improvement
- Most words are seen in large corpus
  - Unknown words: < 5% tokens, < 10% types
  - Simple prefix splitting reduced to half
- Phrase translation captures some of the agreement information
- Local word reordering in the decoder reduces word order problems
- We still believe that morphology could give an additional improvement

Requirements for Improvements

- Data
  - More specific data: We have large corpus (UN) but only small news corpora
  - Manual dictionary could help, it helps for Chinese
- Better use of existing resources
  - Lexicon not trained on all data
  - Treebanks not used
- Continues improvement of models and decoder
  - Recent improvements in decoder (word reordering, overlapping phrases, sentence length model) helped for Arabic
  - Expect improvement from named entities
  - Integrate morphology and alignment