معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute

Member of Qatar Foundation عضو في مؤسسة قطر

The QCN System for Egyptian Arabic to English Machine Translation

Presenter: Hassan Sajjad

QCN Collaboration

Ahmed El Kholy
Wael Salloum

Nizar Habash

Ahmed Abdelali
Nadir Durrani
Francisco Guzmán
Preslav Nakov
Hassan Sajjad
Stephan Vogel

NIST Egyptian Arabic-English Dataset

- Three genres
  - SMS, Chat and CTS
- Dataset distribution
  - Approximately 3000 sentences for tuning
  - The rest is used for training
- Development sets provided by NIST
  - Test: devTest
  - TestG: gold devTest

Baseline System Settings

- Phrase-based SMT system with the following settings:
  - MGIZA for alignment
  - Phrase tables with Kneser-Ney smoothing
  - Lexicalized reordering
  - Operation sequence model
  - Tuning using PRO and MIRA
  - Minimum Bayes risk decoding
  - Cube pruning
  - Other Moses defaults…

Important Modules

Data Preprocessing

- Arabizi to UTF8 conversion
- Normalization
- Speech markups removal
- Cleaning
- Intended vs. literal meaning
- Egyptian Arabic segmentation
- Egyptian Arabic to MSA conversion

System Features

- Class-based models
- Neural network joint model
- Interpolated language model
- Sparse features
- Adaptation
- Unsupervised transliteration model
- System combination

Data Preprocessing

- Arabizi to UTF-8 using **3arrib**

- Normalization
  - Emoticons e.g. ->:=P
    - Tokenizer splits them into single units like - > : = P
    - Normalizing emoticons to their original form
  - Fixed character repetitions on both Arabic and English side
    - Map repetitions like hahahahahah to one from say, haha
    - Convert emphasis repetitions like Yessss to their original form

- Removing markups e.g. %fw, %fp, {laugh}

Data Preprocessing: Egyptian Segmentation

- Segmentation of Egyptian Arabic using **MADAMIRA**
  - ATB, S2, D3

up to +3 BLEU points

	SMS		CHT		CTS	
	Test	TestG	Test	TestG	Test	TestG
No segmentation	21.02	21.64	20.27	22.34	20.60	23.36
D3	23.68	23.41	23.22	25.97	21.72	24.89
S2	23.62	23.66	22.82	25.41	21.61	24.67
ATB	23.57	23.50	22.82	26.01	21.68	24.83

Data Preprocessing: Egyptian to MSA Conversion

- Character-level system to convert Egyptian words to MSA
  - e.g. بيتكلم to يتكلم

Mixed results

	SMS		CHT		CTS	
	Test	TestG	Test	TestG	Test	TestG
Egyptian	21.02	21.64	20.27	22.34	20.60	23.36
Converted MSA	21.54	21.82	20.70	22.77	21.30	23.81
Converted MSA, ATB	21.32	21.06	21.55	23.70	21.73	24.30

- Gains are low compared to the system trained using Egyptian segmentation

- **Highly dialectal nature of the data**
  - requires more lexical substitution than character-level changes

Data Preprocessing: Tuning Dataset Issues

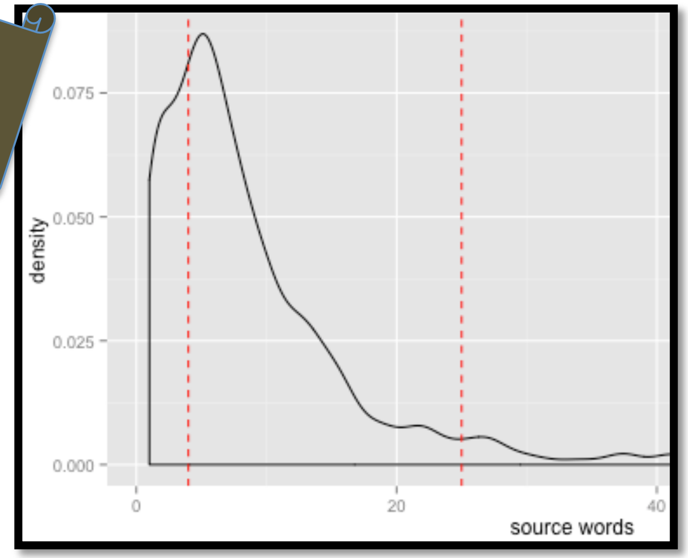
- Missing markers of literal and actual translations in references
- Imbalanced length ratio (i.e. English sentence is 2x of an Arabic sentence)
- Problem: Imbalanced tuning sentences will result in bad tuning weights

Data Preprocessing: Cleaned Tuning Dataset

• Removing sentences with abnormal word length (4<length<25) and length ratio in either source or target

+0.63 on CTS

Genre	Test	[illegible]
SMS - unfiltered tune	23.57	23.56
SMS - filtered tune	23.35	24.36
CTS - unfiltered tune	22.07	25.10
CTS - filtered tune	22.70	25.22



Data Preprocessing: Ex. Noisy References

- Literal meaning is sometimes noisy
- Solution: We used the intended meaning only

Source	احدث الاغاني يا خارجة من باب الارشاد و واخدة قرض من الدوحة، دوحة، دوحة دوحة
Reference	The latest song: [O Muslim Brotherhood who are borrowing money from Doha / O that cute girl who just took a fresh shower, with her cheeks beautifully reddish].
Actual meaning	The best songs oh who you are leaving from the door of Ershad and borrowing money from Doha, Doha, Doha

Important Modules

Data Preprocessing

- Arabizi to UTF8 conversion
- Normalization
- Speech markers removal
- Cleaning
- Intended vs. literal meaning
- Egyptian Arabic segmentation
- Egyptian Arabic to MSA conversion

System Features

- Class-based models
- Neural network joint model
- Interpolated language model
- Sparse features
- Adaptation
- Unsupervised transliteration model
- System combination

System Features: Class-based Models

- Map words into a coarse representation
  - Reduces data sparseness
  - Generalizes the data
- Word clusters using mkcls (k=50, 500)
  - Translation model
  - OSM model over cluster IDs

Consistent improvement up to +0.6 points

	SMS		CHT		CTS	
	Test	TestG	Test	TestG	Test	TestG
Baseline	24.22	24.33	23.02	25.60	21.93	24.88
+ class-based models	**24.63**	**25.16**	23.18	**26.30**	22.20	25.04

System Features: Neural Network Joint Model

- Distributed representation of words
  - Similar to class-based models
    - Reduces data sparseness
    - Generalizes the data

+0.0 on CTS

	SMS Test	SMS TestG	CHT Test	CHT TestG	CTS Test	CTS TestG
Baseline	24.58	24.33	24.02	27.11	22.64	24.95
+ NNJM Model	**25.01**	25.72	24.24	27.41	22.68	25.21

System Features: Genre-Based Interpolated LM

- Divide the available data into groups such as target side of
  - available Egyptian data
  - available Chinese data
  - MSA News
  - MSA non-News
- Minimize the perplexity on each genre's tuning set

up to +1 BLEU points

	SMS		CHT		CTS	
	Test	TestG	Test	TestG	Test	TestG
Concatenated LM	24.19	24.00	23.34	25.89	22.75	25.09
Interpolated LM	**25.20**	**25.04**	**23.48**	**26.16**	**23.01**	**25.67**

System Features: Additional Features

- Domain indicator features
- Source and target word deletion features

Mixed results

	SMS		CHT		CTS	
	Test	TestG	Test	TestG	Test	TestG
Baseline	24.58	24.82	23.36	26.11	22.64	24.95
+ sparse features	24.54	25.36	**24.02**	**27.11**	21.61	24.08

System Features: Adaptation

- Egyptian data with three genres
- MSA data

- Techniques
  - Concatenation
  - Phrase table merging
  - Back-off phrase tables

System Features: Adaptation

- Various combination of available Egyptian data for training
- Testing on SMS genre

Concatenation works the best!

Training	Test	TestG
SMS	21.30	21.99
CAT(SMS, CHT, CTS)	23.78	23.20
SMS, Backoff(CHT,CTS)	22.55	23.00
CAT(SMS,CHT), Backoff(CTS)	22.54	23.20
MergePT(CAT(SMS,CHT),CTS)	23.69	24.40

System Features: Adaptation

- MSA phrase tables – Backoff and Merging
- Helps to translate OOV words which would also help in human evaluation

Merging with MSA translates OOV words

Training	Test	Tes[illegible]
CAT(SMS, CHT, CTS)	23.78	23.20
CAT(SMS, CHT,CTS), Backoff(MSA)	23.70	23.64
MergePT(CAT(SMS,CHT,CTS),MSA)	23.83	23.60

System Features: Unsupervised Transliteration Mining

- Used unsupervised transliteration mining module (implemented in Moses) to transliterate OOV words
  - extracts a list of candidates from parallel training sentences
  - mines transliteration pairs
  - builds a phrase table of transliteration options
  - Post-processes the machine translation output

No improvement

- Most of the OOVs are non-named entities
  - Require translation rather than transliteration

System Combination

- Combine machine translation output of various systems

Drop in CTS

	SMS		CHT			
	Test	TestG	Test	TestG	Test	TestG
Egyptian D3	25.28	26.05	23.87	27.07	**23.34**	**26.05**
Egyptian S2	24.93	25.61	24.09	27.01	22.11	24.50
Egyptian ATB	25.13	25.80	24.24	27.41	22.83	25.56
Egyptian ATB + MSA backoff	25.20	25.04	23.48	26.16	23.01	25.67
Output combination	**26.13**	**26.79**	**24.86**	**27.95**	22.89	25.88

Summary

- Data preprocessing is one of the major challenges in this translation task

- Normalization such as handling emoticons, fixing repetitions and cleaning helps to achieve better alignment

- Improvements of each module vary by genre

- Consistent improvements
  - Egyptian Arabic segmentation (up to +3 points)
  - Genre-based interpolated LM (up to +1 points)
  - Class-based models (up to +0.6 points)

Thank you