FBK @ IWSLT 2009

Nicola Bertoldi, Arianna Bisazza, Mauro Cettolo, Marcello Federico
FBK - Fondazione Bruno Kessler
Via Sommarive 18, 38123 Povo (TN), Italy
Germán Sanchis-Trilles
Instituto Tecnológico de Informática - Universidad Politécnica de Valencia
Camino de Vera s/n, 46022, Valencia, Spain

Main Contributions

- BTEC Arabic-English and Turkish-English:
  - Special effort on **linguistic preprocessing** for morphologically rich source languages
  - In particular **word segmentation** and **lexical approximation** techniques
  - Dealing with mismatch in word granularity between source languages and English
- CT English-Chinese and Chinese-English:
  - Focus on **language model adaptation**
  - Mixture of n-gram language models, obtained by **clustering** training data
  - Mixture weight estimation at the level of single source sentence or complete test set

Linguistic Pre-Processing for Morphologically Rich Languages

- Morphological segmentation of Turkish:
  - vowel harmony (+ other phonological phenomena)
    ⇒ systematic stem and suffix allomorphy
  - agglutinative language
    ⇒ huge variety of possible segmentation schemes

    ⇓
  - tag notation abstracts from suffix allomorphy. Example:
    `elim` → `el+P1sg` ('my hand'), `kolum` → `kol+P1sg` ('my harm')
  - our best segmentation scheme *MS11* handles nominal case, possessive, copula and verb person suffixes:

Turkish:	**Arabic:**
Morph. analysis (Oflazer, 1994)	
Morph. disambiguation (Sak & Saraçlar, 2007)	Morph. segmentation (MADA/AMIRA)
Suffix tags split/removal (tested 11 schemes)	
Lexical approximation	Lexical approximation

Preprocessing pipelines

odamdayım	→ *oda/m/da//yım*
('I am in my room')	`oda+Noun+A3sg/+P1sg/+Loc/^DB+Verb+Zero+Pres/+A1sg`
bayanın çantasını gördüm	→ *bayanın çantasını gör/düm*
('I saw the lady's bag')	`bayan+Noun+A3sg+Pnon+Gen çanta+Noun+A3sg+P3sg+Acc gör+Verb+Pos+Past/+A1sg`

- Morphological segmentation of Arabic:
  - specific tokenization *(arTok)*: removal of short vowels and normalization of UTF8 characters and digits
  - comparison of two state-of-the-art segmenters: MADA and AMIRA

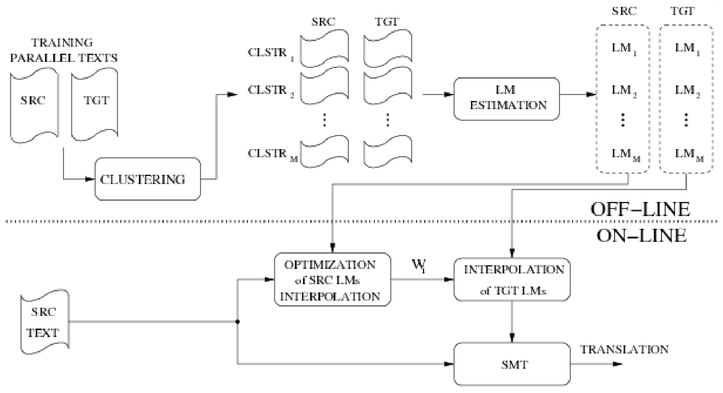
		Example: 'And she will say it to her colleague.'
Baseline:		`wstqwlh lzmylhA` `[and-she-will-say-it] [to-her-colleague]`
MADA:	heavy-weight, based on	`w+ s+ tqwlh l+ zmylhA`
(Habash & Rambow, 2005)	morphological analysis	`[and] [will] [she-say-it] [to] [her-colleague]`
AMIRA:	light-weight, based on	`w+ stqwl +h l+ zmyl +hA`
(Diab & al., 2004)	5-characters context	`[and] [she-will-say] [it] [to] [colleague] [her]`

- Lexical approximation:
  replace OOV words in the test with morphologically similar words of the training
  - deterministic choice of 1-best replacer
  - Turkish: choose word sharing lemma and largest number of suffix tags
    Example: `çıkışlar (çık+Verb+Pos^DB+Noun+Inf3+A3pl)` → `çıkış (çık+Verb+Pos^DB+Noun+Inf3+A3sg)`
  - Arabic: progressively remove prefix and suffixes from the OOV word until a replacer is found.
    Example: `tmddy` → `tmdd` → `mdd`, `qmySk` → `qmyS`

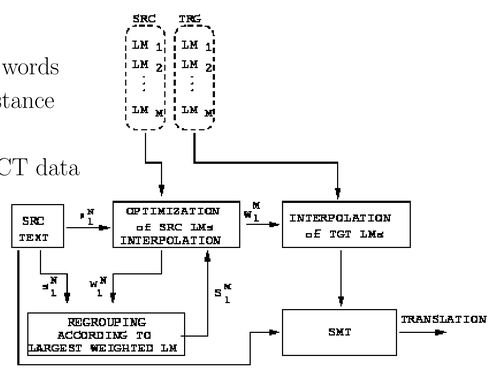
Online Language Model Adaptation for Spoken Dialog Translation

- Model adaptation

  LM score is given by either single LM (`baseline`) or mixture of (smaller) LMs: $p(\mathbf{e}) = \sum_{i=1}^{M} w_i p_i(\mathbf{e})$



- Clustering using dialog annotations:
  - Each dialog is represented as a bag of both source and target words
  - CLUTO package was employed: `direct` clustering, cosine distance
  - 2, 4, 6 and 8 clusters
  - One set of LMs for each cluster + additional LM on BTEC+CT data
- On-line weight optimization:
  - Set specific weights (over complete source side of test set)
  - Sentence specific weights (one set of weights for each source sentence)
  - Two-step weight optimization: See figure.



Evaluation Results

- Baseline : standard setup for Moses SMT toolkit
- BTEC Turkish-English
  - Best segment. scheme *(MS11)* dramatically lowers test's OOV, minimizes differences in word granularity between TR and EN, reduces training dictionary size and data sparseness.
  - MERT on devset1 using gold reference only
  - Distortion limit (DL) set to 10, due to high word order mismatch
  - Morph. segmentation yields 5 points BLEU improvement
  - Lexical approx. does not improve in *-drop-unknown* conditions
  - Unlimited distortion results inconsistent across test sets

	training		devset2	
preprocessing	|W|	|V|	OOV%	H(bits)
-	139,514	17,619	6.16	59,435
MS11	168,135	10,450	2.54	57,379

Effect of preprocesssing on Turkish data

system	MS11	lex.appr.	DL	devset2	test
baseline	-	-	10	54.80	51.82
primary	+	-	10	**59.77**	56.77
contrastive1	+	+	10	59.24	56.75
contrastive2	+	-	∞	59.02	**57.04**

BTEC Turkish-English results (%BLEU)

system	prepr.	lex.appr.	LM	devset7	test
baseline0	-	-	msb	51.87	51.36
baseline1	arTok	-	msb	52.38	51.75
primary	mada	-	msb	**54.68**	52.23
contrastive1	mada	+	msb	54.52	52.92
contrastive2	amira	-	msb	54.60	**53.36**
contrastive3	mada	-	kn	53.78	51.92

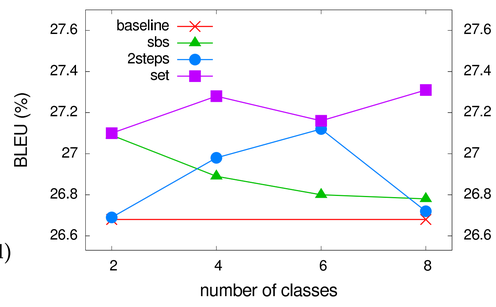
BTEC Arabic-English results (%BLEU)

- BTEC Arabic-English
  - Training data: train + devsets 2, 3 and 6 (with gold reference only)
  - MERT on devset1 using all references
  - Specific tokenization alone yields around half point BLEU improvement (51.36 to 51.75 on test)
  - Morph. segmentation through MADA yields additional 2.3 points on devset7, but only 0.5 on test
  - AMIRA results inconsistent across test sets
  - Lexical approx. results also discrepant: improvement only on the official test

- CT English-Chinese
  - Development set of CT task used for MERT, then included into training corpus
  - Development sets of previous campaigns not included, only their vocabulary
  - Same system (differently tuned) for ASR and CRR
  - Improvements in terms of perplexity are only partially mirrored into translation quality
  - Primary run: six dialog clusters, 2step weight estimation

		ASR		CRR	
system	submission	BLEU	p/r	BLEU	p/r
baseline	contrastive3	32.75	61.7/59.1	40.40	68.4/66.3
2steps	primary	33.37	63.2/59.4	40.05	68.5/66.1
set	contrastive1	33.71	63.3/59.6	40.33	68.8/66.2
sbs	contrastive2	33.20	63.2/59.5	39.73	68.1/65.7

CT English-Chinese results (%BLEU and precision/recall)



- CT Chinese-English
  - Same setup as for English-to-Chinese

		ASR		CRR	
system	submission	BLEU	p/r	BLEU	p/r
baseline	contrastive3	30.01	63.3/63.2	31.82	66.4/67.3
2steps	primary	30.13	63.5/63.4	31.92	66.5/67.8
set	contrastive1	29.92	63.6/62.7	32.15	66.5/67.6
sbs	contrastive2	29.96	64.0/63.6	31.87	66.7/67.6

CT Chinese-English results (%BLEU and precision/recall)

Summary and Future Work

- Specific **linguistic preprocessing** is crucial for morphologically rich languages
- TODO: refine our Turkish segmentation schemes by addressing verbal suffixation in a better way
- TODO: feed Moses with multiple options for lexical approximation

- **Adaptation** yields limited gains in BLEU
- Observed big gains in perplexity → room for improvement
- TODO: address larger tasks, involving unsupervised clustering and source-to-target weight map

Acknowledgements

This work was supported by the EuroMatrixPlus project (IST-231720), which is funded by the European Commission under the Seventh Framework Programme for Research and Technological Development, and by the Spanish MEC under scholarship AP2005-4023 and grant CONSOLIDER Ingenio-2010 CSD2007-00018.