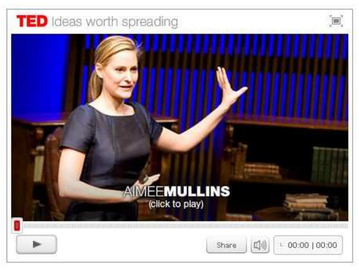
FBK @ IWSLT 2010

Arianna Bisazza, Ioannis Klasinas*, Mauro Cettolo, Marcello Federico

FBK - Fondazione Bruno Kessler - Trento, Italy
*IRIT - Université Paul Sabatier - Toulouse, France

TALK

– Moses baseline trained on all data
– In-domain **data selection** using perplexity
– Parallel data filtering by **fragment extraction**



BTEC

– New morphological **segmentation rules** for Turkish
– Combine different segmentation schemes into **lattice input**
– Arabic: Use of additional resources with multiple phrase table decoding

Data Selection Techniques

Available data: - small in-domain corpus TED,
- huge generic corpus Giga

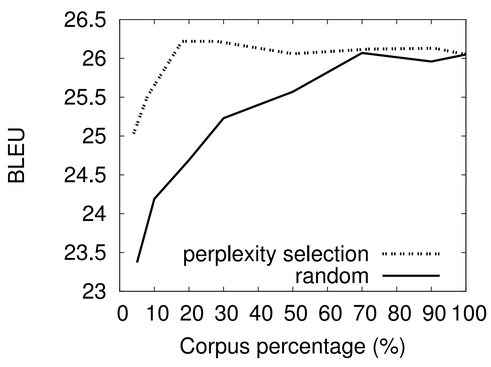
Baseline: log-linear interpolation of PhTables and LMs trained on all data (ALL)

Data selection techniques for Giga data:

- long sentence filtering (ALLflt)
- in-domain sentence selection by perplexity (PPsel) outperforms ALL using only 30% of data

Corpus	Sentences	Tokens EN	Tokens FR
TED	84k	0.85M	0.89M
ALL	31.5M	920M	1062M
ALLflt	24.6M	558M	633M
PPsel	9.5M	267M	303M
FRAG	–	126M	140M

TALK task corpora statistics



SYSTEM PhTable	SYSTEM LM	BLEU
TED	TED	24.44
TED	TED,ALL	26.65
ALLflt	TED,ALL	28.61
TED,ALLflt	TED,ALL	29.51
TED,FRAG	TED,ALL	29.03
TED,ALLflt	TED,PPsel(30%)	29.75
TED,PPsel(30%)	TED,PPsel(30%)	**29.92**

%BLEU scores on dev set (tokenized, no_case)

- parallel fragments extraction (Cettolo & al. 2010): minor loss (-0.48%) wrt ALL using only 15% of data

SMT Model Coverage and Rich Morphologies

Turkish & Arabic: rich morphologies ⇒ negative impact on phrase-based MT for small tasks

Morphological segmentation of Turkish

- Selectively split/remove suffixes from morph. analysed data (Bisazza & Federico 2009):

```
TR:     öksürüğümü durduramıyorum
morph:  öksürük+P1sg+Acc  dur+Caus+Able+Neg+Prog1+A1sg
MS11:   öksürük  +P1sg  dur+Caus+Able+Neg+Prog1   +A1sg
MS13:   öksürük  +P1sg  dur+Caus+Able+Prog1   +Neg  +A1sg
MS14:   öksürük  +P1sg  dur+Caus+Prog1  +Able  +Neg  +A1sg
MS15:   öksürük  +P1sg  dur+Prog1  +Caus  +Able  +Neg  +A1sg
        [cough]   [my]   [stop]   [make]  [can]  [not]  [I]
EN:     I can't stop coughing (lit.  I cannot make my cough stop)
```

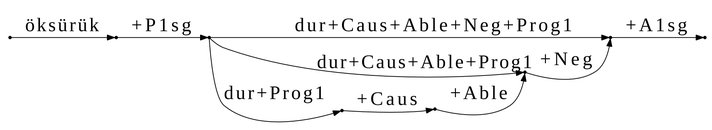
- Baseline Morphological Segmentation scheme ("MS11") deals with:

nominal cases, possessive suffixes, few rules for verbs (person subj. & copula)

- + New rules for verbal inflection:
  - ◇ **negation** (MS13): after applying MS11, extract suffix *-mA* from the verb and split
  - ◇ **ability** (MS14): extract suffix *-Abil* (English 'can') from the verb and split
  - ◇ **voice suffixes** (MS15): extract passive and causative suffixes from the verb and split

Segmentation lattice input

- Hard to find optimal segmentation rule set given task & language-pair
- Possible solution: combine various degrees of segmentation in input
  ⇒ decoder can choose word-level-optimal segmentation path
- Training set = differently segmented versions of train, concatenated
- Example lattice combining MS11 + MS13 + MS15:



TR: öksürüğümü durduramıyorum
(EN: I cannot make my cough stop)

segmentation	BLEU – NIST
MS11	60.30 – 9.367
MS13	58.98 – 9.357
MS14	57.76 – 9.373
MS15	60.32 – 9.575
MS(11+13+15)lattice	**60.41 – 9.650**

%BLEU–NIST scores on dev2: segm. schemes & lattice combination

Arabic: Multiple PhTable decoding

- Still many OOVs in Arabic (4.10% on dev7) after morphological segmentation with AMIRA
- Add PhTable trained on large *out-of-domain* corpus: news from NIST-MT09, 6.2M words
- Small in-domain PhTable ⋃ large news PhTable ⇒ double weight set, MERT-optimized

Official Results

- ASR input: translation of 1-best hypotheses
  – no optimization for this condition
  ⇒ drop in performance :-(
- Overall best results using 30% of Giga selected by perplexity

Text format	Dev(%BLEU)
tokenized, no_case	29.98
tokenized, case	28.47
detokenized, case	27.07

Postprocessing impact

	SYSTEM	Dev set BLEU	Dev set TER	Test set BLEU	Test set TER
Correct text	Primary	27.07	0.5732	29.90	0.5350
	Contrastive	26.65	0.5781	28.67	0.5436
ASR	Primary	13.18	0.7386	15.19	0.6980
	Contrastive	13.19	0.7403	14.66	0.7022

TALK task: official FBK scores

```
Src     Now, a language is not just a body of vocabulary
TED     Maintenant, un langage n'est pas juste un corps de vocabulaire
PPsel   Maintenant, une langue n'est pas seulement un corps de vocabulaire
Ref:    De nos jours, une langue n'est pas uniquement un ensemble de vocabulaire
```

Translation example

Official Results

%BLEU–NIST scores on dev and test sets:

Turkish-English

Segm.lattice better than plain input on dev2 and test2010 but simple MS15 wins on test2009
[distortion limit =10]

Run	segmentation	dev2	test2009	test2010
primary	MS11+13+15	**60.41 – 9.650**	57.70 – 8.612	**53.29 – 8.443**
contr1	MS15	60.32 – 9.575	**58.28 – 8.660**	52.46 – 8.441
contr3	MS11	60.30 – 9.367	57.21 – 8.422	52.14 – 8.136

Arabic-English

Good reduction of OOV (4.10% to 2.71% on dev7) with 2 ph.tables, but no consistent improvement on translation quality
[distortion limit = 6]

Run	ph.tables	dev7	test2009	test2010
primary	btec	**55.02 – 8.735**	**52.04** – 7.494	43.07 – 7.254
contr.	btec+news	54.20 – 8.620	51.08 – **7.514**	**43.76 – 7.255**

Acknowledgements

- EuroMatrixPlus project (IST-231720)
- French Allocation de Recherche, contract 26076-2007