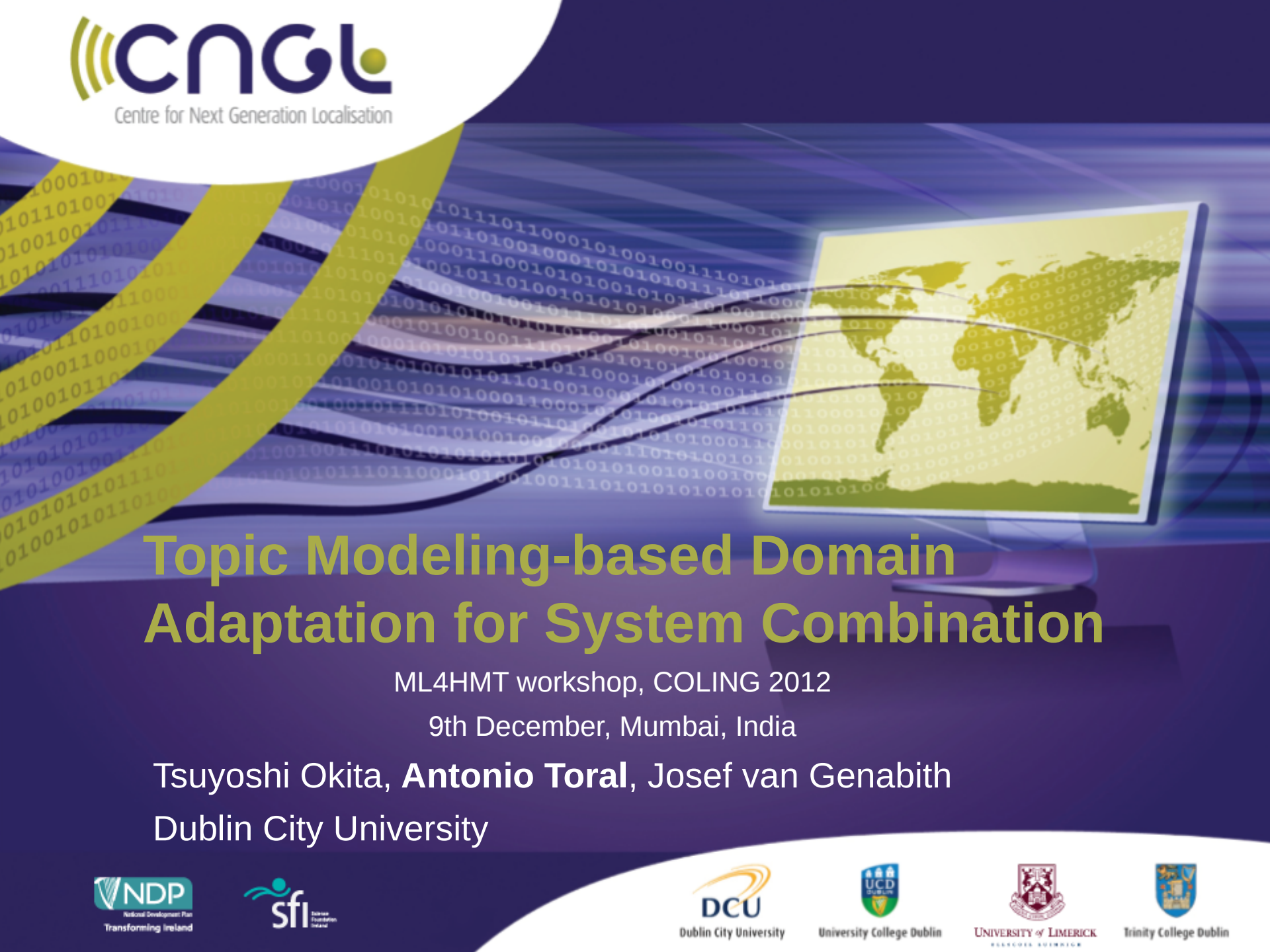

Topic Modeling-based Domain Adaptation for System Combination

ML4HMT workshop, COLING 2012

9th December, Mumbai, India

Tsuyoshi Okita, **Antonio Toral**, Josef van Genabith

Dublin City University

Centre for Next Generation Localisation

Contents

- Intro
- Method
- Results
- Conclusions
- Future work

Intro

- Extension to DCU system combination modules: domain adaptation
- Participation in system combination task

Intro

- Background idea: genre classification of training data
  - Most approaches supervised
  - Cache-based approach (Tiedemann, 2010) does not need notion of genre
- Idea: employ unsupervised document classification to cluster the documents

Intro

- Hypothesis: genre of test and tuning sets exhibit variance, hence out-of-domain effects
- This causes variance in performance of MT system
- Methods explored:
  - Identify and remove out-of-domain data from tuning
  - Train on in-domain partitioned data

Method. System combination module

- Two step system combination
  - Set parameters on tuning (MERT)
  - Use these parameters to decode test
- Other features
  - MBR decoding, BLEU as loss function
  - TERp as alignment metrics in monolingual word alignment

Method. Document classification

- Latent Dirichlet Allocation (LDA)
  - Topics as multinomial distributions over word-types in the corpus
  - Documents as a mixture of topics
  - Classifies documents into given number of classes

Method. Document classification

- Out-of-domain data cleaning from tuning set
  - Fix number of classes [500, 1000]
  - LDA on tuning and test sets
  - Detect classes that contain data only from tuning set
  - Discard corresponding sentence pairs from tuning set

Method. Document classification

- In-domain data partitioning
  - Fix number of classes [1, 5]
  - LDA on tuning and test sets
  - Separate each class of tuning and test (keep original and new indexes)
  - Run system combination on each class
  - Reconstruct system combined results preserving original index

Evaluation. Setting

- ML4HMT-2012 task
  - Output of 4 MT systems
    - 2 RBMT: Apertium (s1), Lucy (s2)
    - 2 SMT: PB Moses (s3), HPB Moses (s4)
  - Data
    - Tuning: 20,000 sentences
    - Test: 3,003

Evaluation. Results (LDA)

	tuning					test				
class 1	20000					3003				
class 2	10213	9787				1821	1182			
class 3	6752	6428	6820			838	962	1203		
class 4	4461	4766	5954	4819		785	432	776	1010	
class 5	3846	3669	3665	3978	4842	542	343	1311	404	403

Evaluation. Results, out-of-domain cleaning

- Process removed 2,207 sentences from tuning set, 11%
- 1 BLEU point loss over baseline system combination

	NIST	BLEU	METEOR	WER	PER
cleaned	7.4945	0.2500	0.5499287	56.6991	42.3032
wo cleaning	7.6846	0.2600	0.5643944	56.2368	41.5399

Evaluation. Results, in-domain partitioning

- Gain 0.33 BLEU over baseline system combination

	NIST	BLEU	METEOR	WER	PER
single best results					
s1	6.4996	0.2248	0.5458641	64.2452	49.9806
s2	6.9281	0.2500	0.5853446	62.9194	48.0065
s3	7.4022	0.2446	0.5544660	58.0752	44.0221
s4	7.2100	0.2531	0.5596933	59.3930	44.5230
topic modeling (testset)					
2 class	7.7036	0.2620	0.5626187	55.8092	41.7783
3 class	7.7134	0.2628	0.5645200	55.8865	41.7171
4 class	7.7146	0.2633	0.5647685	55.8612	41.7264
5 class	7.6245	0.2592	0.5620755	56.9575	42.6229
system combination without topic modeling					
syscom	7.6846	0.2600	0.5643944	56.2368	41.5399

Conclusions

- Contribution: domain adaptation to system combination via unsupervised document clustering (topic modelling)
- Results
  - Out-of-domain cleaning: lost 1 BLEU point compared to baseline system combination
  - In-domain partitioning: gain 0.33 BLEU over baselines system combination. 1.02 BLEU over best MT system

Future work

- Explore this topic further
  - Use larger datasets
  - Explore bigger values for classes in partitioning (max here 5)
- Other ideas for system combination
  - Correction of output based on corresponding tokens and PoS tags from the source and target, ~Automatic Post Editing

End

Thanks for your attention!

आभार

Topic Modeling-based Domain Adaptation for System Combination

ML4HMT workshop, COLING 2012

9th December, Mumbai, India

Tsuyoshi Okita, **Antonio Toral**, Josef van Genabith

Dublin City University