Evaluating Syntax-Driven Approaches to Phrase Extraction for Machine Translation

Ankit **S**rivastava	**S**ergio **P**enkale	**D**eclan **G**roves	**J**ohn **T**insley
CNGL, DCU	*CNGL, DCU*	*Traslán*	*NCLT, DCU*

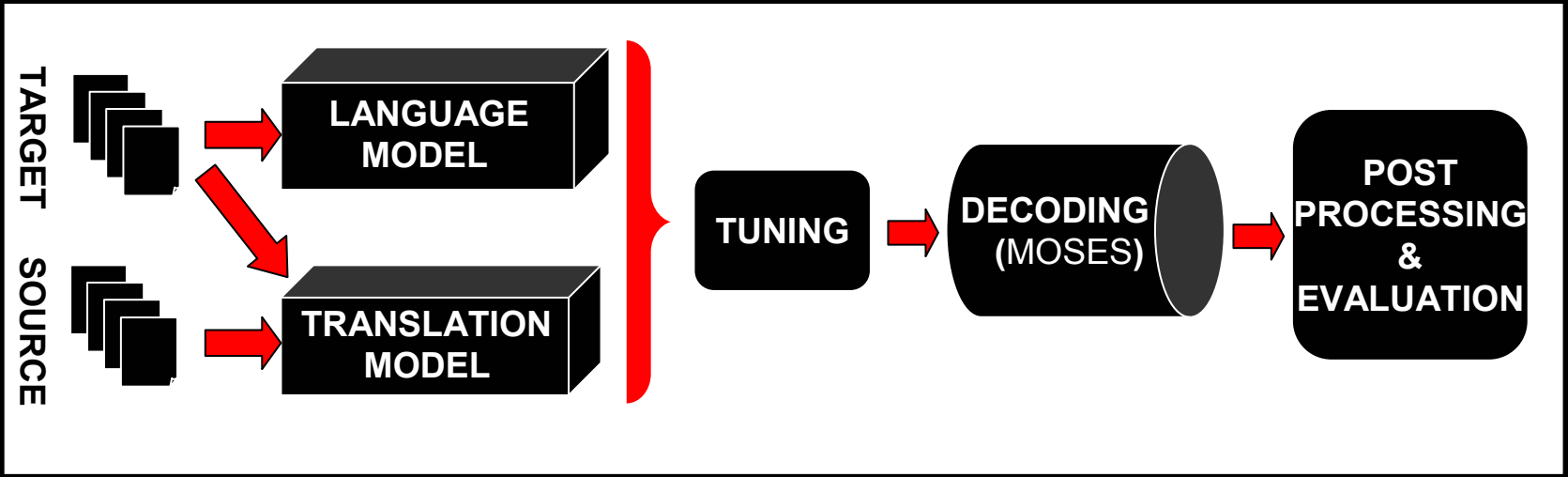
3rd International Workshop on Example-Based Machine Translation, 12-13 November, 2009: Dublin, Ireland

about

- EBMT-influenced data sources used in a PB-SMT model (Moses)
  - Marker Hypothesis
  - Parallel Treebanks
- Lessons learned from work carried out over a number of years at DCU
- Focus on techniques for supplementing Moses phrases with syntactically motivated phrases

pb-smt system

- Moses framework [**Koehn et al., 2007**]

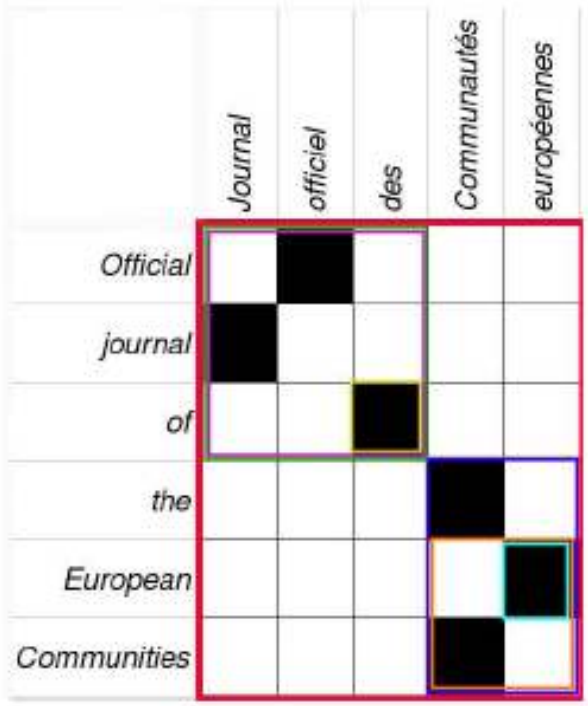


- Translation model
  - Heuristics-based phrase extraction from bidirectional word alignments
  - Syntactically-motivated phrase extraction: marker / treebank

Moses phrases: an example

Official journal of the European Communities
Journal officiel des Communautés européennes

Official journal	↔	Journal officiel
Official journal of	↔	Journal officiel des
Official journal of the \ European Communities	↔	Journal officiel des \ Communautés européennes
of	↔	des
of the European Communities	↔	des Communautés européennes
the European Communities	↔	Communautés européennes
European	↔	européennes



marker-based

- Chunk sentences on encountering a ‘marker’ word
  - Founded on the Marker Hypothesis [**Green, 1979**]
  - Marker words are closed class of lexemes / morphemes
  - Each marker word associated with a marker category (tag)
  - 7 marker categories identified. E.g. DET, PREP, PRON
  - Each marker chunk must contain at least 1 non-marker word
- Align bilingual marker chunks
  - Use marker tag and relative positions in the sentence
  - Use cognate and MI scores
- Obtain marker-based phrase pairs

Marker-based: an example

That is almost a personal record for me this autumn
C' est pratiquement un record personnel pour moi cet automne

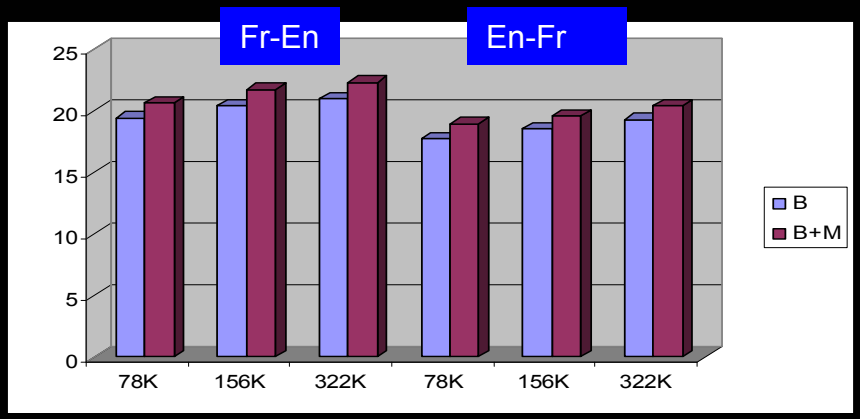
<DET>That is almost **<DET>a** personal record **<PREP>for <PRON>me <DET>this** autumn
<DET>C' est pratiquement **<DET>un** record personnel **<PREP>pour <PRON>moi**
<DET>cet automne

<DET>That is almost **<DET>**a personal record **<PREP>**for me this autumn
<DET>C' est pratiquement **<DET>**un record personnel **<PREP>**pour moi cet automne

That is almost ↔ C' est pratiquement
a personal record ↔ un record personnel
for me this autumn ↔ pour moi cet automne

marker-based: direct

- Merging phrase pairs in a single phrase table
- Fr-En Europarl data: (3-gram lang model, Pharaoh decoder)
- System performance as training data increases
- 13% new phrases added via marker-based phrases



- Es-En Europarl data: (200K train, 5-gram lang model, Moses decoder)

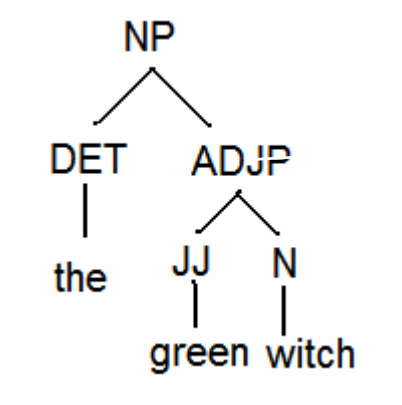
System	BLEU	NIST	METEOR
Baseline	0.3079	.5590	0.6025
1-count	0.3078	7.5775	0.6024

treebank-based

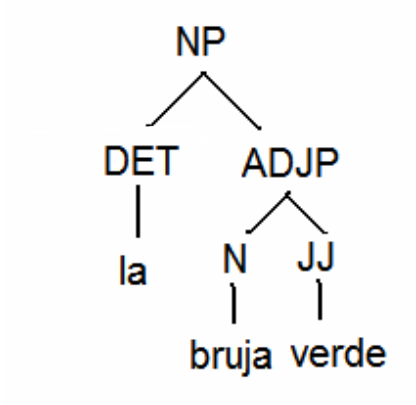
- Monolingual parsing of sentences
  - Parse both sides
  - Requires constituency-structure parsers
- Align bilingual parse trees
  - Requires a sub-tree aligner [**Zhechev & Way, 2008**]
- Get aligned phrases
  - Extract surface-level chunks
- Also implemented using dependency structure
  - Using off-the-shelf dependency parsers
  - Head percolation of constituency trees [**Magerman, 1995**]

treebank-based: an example [con]

the green witch



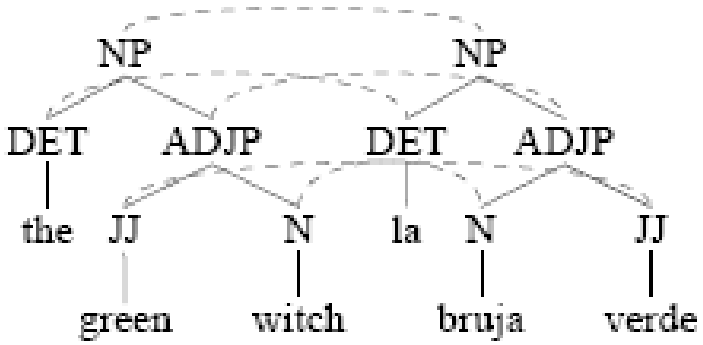
la bruja verde



treebank-based: an example [con]

the green witch

la bruja verde



the green witch	↔	la bruja verde
green witch	↔	bruja verde
the	↔	la
green	↔	verde
witch	↔	bruja

treebank-based: direct

- En-Es Europarl data: (700K train, 5-gram lang model, Moses decoder)
- Moses (Baseline), Constituency (Syntax)
- Merging phrase pairs in a single phrase table
- 24M phrases in Baseline *Vs* 6M phrases in Syntax
- 4.87% overlap between Moses and Syntax
- 16.79% new phrases added

System	BLEU	NIST	METEOR
Baseline	0.3341	7.0765	0.5739
+Syntax	**0.3397**	**7.0891**	0.5782
Syntax_only	0.3153	6.8187	0.5598

treebank-based: direct

- Fr-En Europarl data: (100K train, 5-gram lang model, Moses decoder)
- Moses (B), Constituency (C), Dependency (D), Percolated (P)
- Merging phrase pairs in a single phrase table (1 / 2 / 3 / 4)
- Compare sizes of B with C/D/P
- Overlap between tables

SMT + Syntax				*Pure Syntax*			
System	**BLEU**	**NIST**	**METEOR**	**System**	**BLEU**	**NIST**	**METEOR**
BASELINE(B)	0.2850	7.00	0.5783				
B + C	**0.2950**	**7.10**	**0.5855**	CON(C)	0.2564	6.55	0.5526
B + D	0.2930	7.08	0.5843	DEP(D)	0.2524	6.59	0.5465
B + P	0.2945	**7.10**	0.5854	PERC(P)	0.2587	6.59	0.5563
B + C + D	0.2929	7.09	0.5848	C + D	0.2632	6.69	0.5556
B + C + P	0.2949	**7.10**	0.5850	C + P	0.2637	6.62	0.5605
B + D + P	0.2939	7.09	0.5849	D + P	0.2657	6.74	0.5583
B + C + D + P	0.2940	7.09	0.5849	C + D + P	0.2690	6.75	0.5614

recap

- Baseline system (Moses):
  - ***source_phrase ||| target_phrase ||| [feature_value]***
- Alternate phrase pairs
  - Marker-based: ***src ||| tgt***
  - Treebank-based (con, dep): ***src ||| tgt***
- Experiments on direct combination
  - Merging phrase pairs and re-estimating probabilities

Other ways to supplement the Moses phrase table with alternate phrase segmentation approaches

Combining strategies

- Direct combination 
- Weighted combination
- Prioritised combination
- Feature-based
- System combination

Weighted combination

- Instead of simple merging, add ‘n’ copies of a type of phrase pair
- This modifies the relative frequency of the syntax-based phrase pairs
- Generally does not improve over direct combination
- Experiments on adding n copies of marker-based phrases
- Experiments on adding n copies of constituency-based phrases

System	BLEU	NIST	METEOR	EBMT%
Baseline	0.3079	7.5590	0.6025	24.21
1-count	0.3078	7.5775	0.6024	23.47
2-count	0.3076	7.5582	0.6020	23.64
4-count	0.3071	7.5609	0.6015	24.34
8-count	0.3083	7.5969	0.6018	26.64
16-count	0.3042	7.5386	0.5986	29.71

System	BLEU	NIST	METEOR
Baseline+Syntax	0.3397	7.0891	0.5782
+Syntax x2	0.3386	7.0813	0.5776
+Syntax x3	0.3361	7.0584	0.5756
+Syntax x5	0.3377	7.0829	0.5771
Half-weights	**0.3404**	**7.1050**	**0.5792**

prioritized combination

- A phrase pair consists of
  src ||| ***tgt*** ||| ***[feature_value]***
- Alternative to direct combination
  - Prioritize set A over set B
  - Add only those B phrase pairs when src not in A
- Experiments on baseline & constituency
  - No improvements over direct combination

System	BLEU	NIST	METEOR
Baseline	0.3341	7.0765	0.5739
+Syntax	**0.3397**	**7.0891**	0.5782
Syntax_only	0.3153	6.8187	0.5598
Syntax Prioritised	0.3339	6.9887	0.5723
Baseline Prioritised	0.3381	7.0835	**0.5789**

feature-based combination

- A phrase pair consists of
  src ||| tgt ||| [feature_value]
- Add a new feature
  - Binary: type of phrase pair
  - MERT tuning assigns weight like other features
- Merging like direct combination
- Experiments on marker-based
  - Improvements in translation quality

System	BLEU	NIST	METEOR	EBMT%
Baseline	0.3079	7.5590	0.6025	24.21
1-count	0.3078	7.5775	0.6024	23.47
Feature	0.3111	7.6004	0.6055	35.09

System combination

- So far, all methods have altered how phrases merged into one phrase table
- An alternative is to combine translated sentences (after decoding) rather than phrase pairs (during training)
- Use MBR-CN system combination [**Du et al., 2009**]
  - Experiments on B/C/D/P
  - Output sentences are unique enough to profit
  - 7.16% relative (4 systems) , 12.3% relative (15 systems)

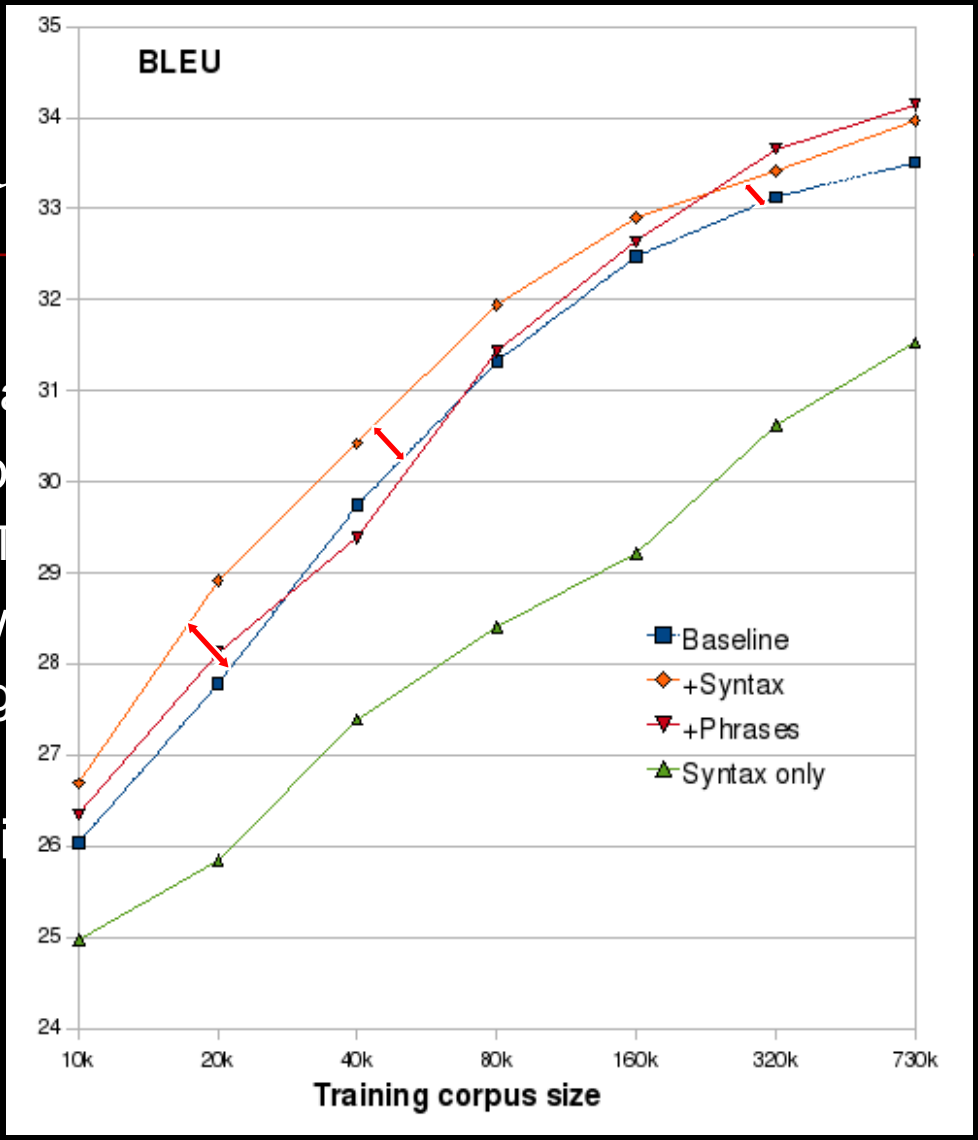
System	BLEU	NIST	METEOR
MBR (4 systems)	0.2952	6.85	0.5784
CN (4 systems)	0.3070	7.06	0.5852
MBR (15 systems)	**0.3260**	7.32	**0.6050**
CN (15 systems)	0.3251	**7.33**	0.6039

lessons learned

- Syntax-based phase pairs are a unique knowledge source
  - Overlap between phrase pairs
- Using only syntax-based phrases deteriorates
  - Large coverage of PB-SMT method
- Supplementing PB-SMT with syntax-based helps
  - Explored 5 different strategies for combining
  - System combination helps the most
- Decrease in gains as training data increases

lessons learned

- Syntax-based phase pairs
  - Overlap between phrase pa
- Using only syntax-based p
  - Large coverage of PB-SMT
- Supplementing PB-SMT w
  - Explored 5 different strateg
  - System combination helps
- Decrease in gains as traini



endnote

- Examined a number of different phrase segmentation approaches for MT
- Explored ways of using linguistic information (borrowed from EBMT research) in a PB-SMT system
- Level of improvement is dependent on amount of training data

- Useful for languages with limited training data and MT systems with a smaller footprint
- Difficult to improve the PB-SMT alignment / extraction / decoding pipeline without significant remodeling

thank you!

- Questions?

- Contact info
  - Declan: dgroves @ traslan.ie
  - Sergio: spenkale @ computing.dcu.ie
  - John: jtinsley @ computing.dcu.ie
  - Ankit: asrivastava @ computing.dcu.ie

Bonus Slide: Sample Output

La commission entend –elle garantir plus de transparence à cet égard ?

- REF: Does the commission intend to seek more transparency in this area?

- MOSES: Will the commission ensure that more than transparency in this respect?
- CON: The commission will the commission ensure greater transparency in this respect?
- DEP: The commission will the commission ensure greater transparency in this respect?
- PERC: Does the commission intend to ensure greater transparency in this regard?