Advancements in Arabic-to-English Hierarchical Machine Translation

Matthias Huck, David Vilar, Daniel Stein, Hermann Ney

EAMT 2011
Leuven, Belgium – May 31, 2011

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University, Germany

Outline

1. Review: Hierarchical phrase-based translation
2. Extensions
   - Shallow rules
   - IBM-style reorderings
   - Soft syntactic labels
   - Lightly-supervised training
   - Discriminative word lexicon
3. Experimental results NIST Arabic→English

Review: Hierarchical Phrase-based Translation

- **Allow for *gaps* in the phrases**
- **Formalization as a *synchronous context-free grammar***
  - **Rules of the form $X \rightarrow \langle \gamma, \alpha, \sim \rangle$, where:**
    - **X is a non-terminal**
    - **γ and α are strings of terminals and non-terminals**
    - **$\sim$ is a one-to-one correspondence between the non-terminals of α and γ**
- ***Parsing-based decoding* (extension of CYK algorithm)**

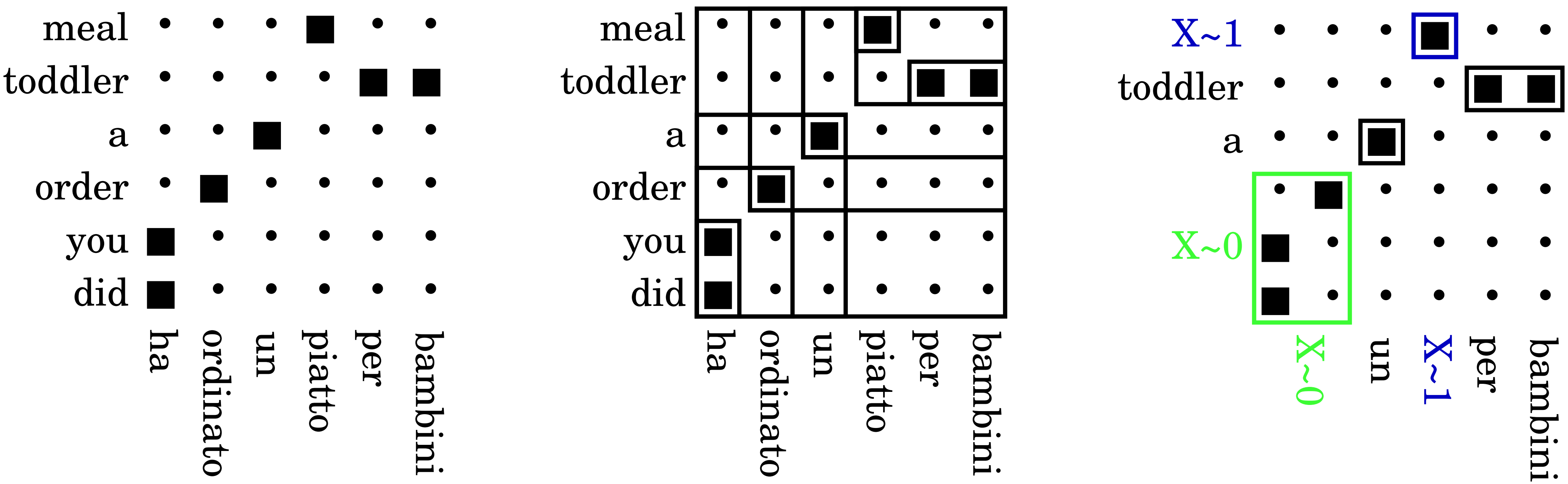
Review: Hierarchical Extraction Process

- **Basic idea:**
  - **Extract standard phrases**
  - **If the extracted phrases contain further sub-phrases, create "holes"**
  - **Assign probabilities using relative frequencies**
- **Main restrictions:**
  - **Maximum of two non-terminals per rule**
  - **Non-terminals must be non-adjacent in the source side**
  - **Rules must have at least one terminal symbol**
- **Additionally: *Initial and glue rule***

$$
\begin{aligned}
S &\rightarrow \langle X^{\sim 0}, X^{\sim 0} \rangle \\
S &\rightarrow \langle S^{\sim 0} X^{\sim 1}, S^{\sim 0} X^{\sim 1} \rangle
\end{aligned}
$$

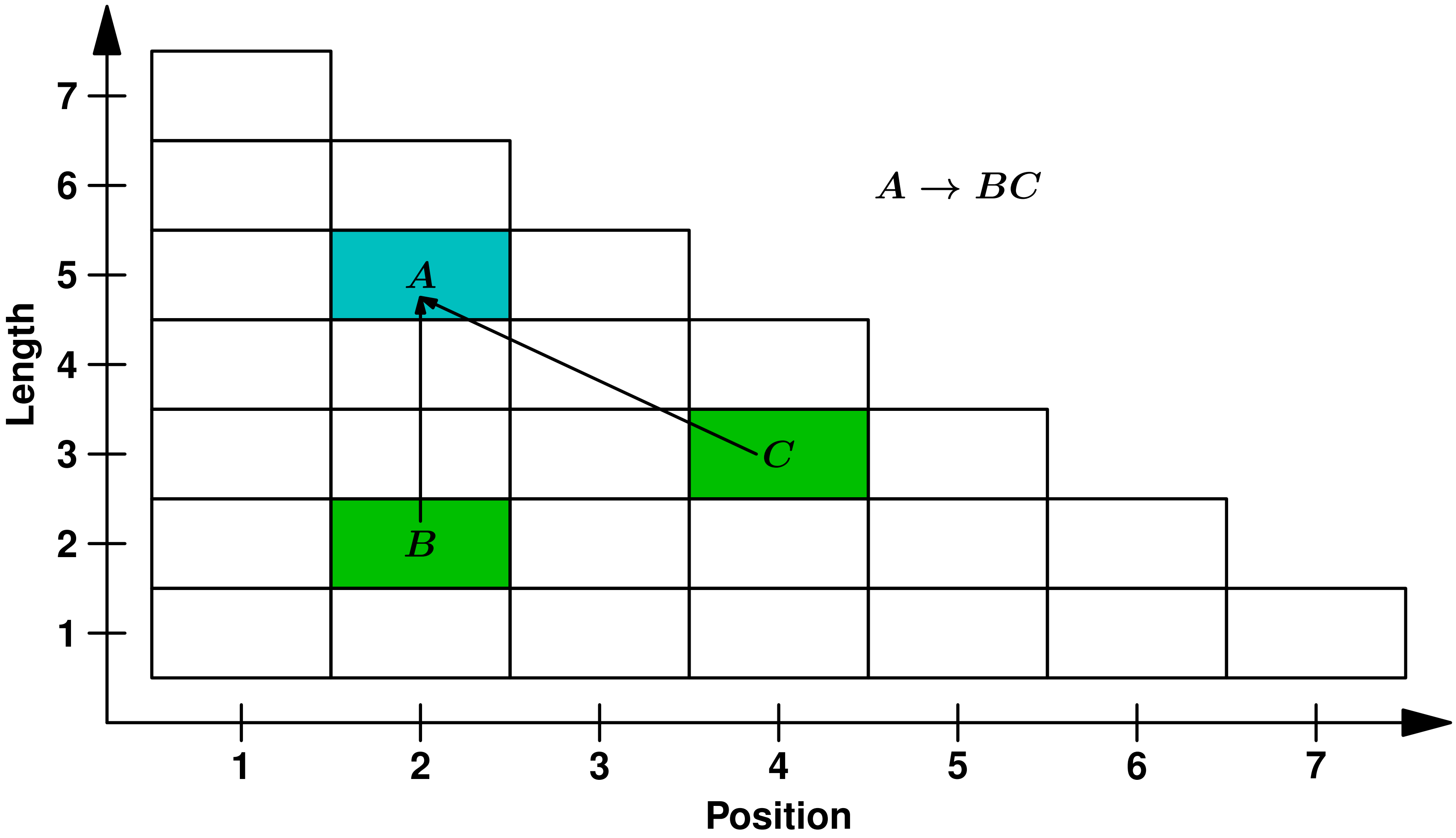
- **Only one *generic non-terminal symbol* X plus the start symbol S**

Hierarchical Rules: Example



Alignment **Standard phrases** **Hierarchical rule**

Review: CYK Algorithm



- Parse tree of the source sentence induces a parse tree of the target sentence
- Additionally to parsing algorithm: Handle translation alternatives
- Cube pruning [Huang and Chiang, ACL 2007]

Hierarchical Phrase-based Translation System

Extensions described here have been integrated into an open source toolkit:

```
                ('-.            .-') _     ('-.
               ( OO ).-.       ( OO ) )_(  OO)
       ,--.   / . --. /,--./ ,--,'(,------.
   .-')| ,|   | \-.  \ |   \ |  |\ |  .---'
  ( OO |(_|.-'-'  |  ||    \|  | )|  |
  | `-'|  |  \| |_.'  ||  .     |/(|  '--.
  ,--. |  |   |  .-.  ||  |\    |  |  .--'
  |  '-'  /   |  | |  ||  | \   |  |  `---.
   `-----'    `--' `--'`--'  `--'  `------'
```

- ***RWTH's open source hierarchical phrase-based translation toolkit*** **(free for non-commercial purposes)**
- **Implemented in C++**
- **See *[Vilar et al., WMT 2010]***
- **`http://www.hltpr.rwth-aachen.de/jane`**

Arabic→English NIST Task

- Our baseline setup: 2.5M sentences of parallel training data
- Systems tuned towards BLEU on MT06
- Results reported on MT08 (news wire and web text) as unseen test set (45K running words)

	Arabic → English (MT08)	
	BLEU [%]	TER [%]
HPBT Baseline	*44.3* ±1.1	*50.0* ±0.9

- Tiny numbers: 95% confidence interval
- For comparison: RWTH's standard *PBT* baseline system (without extensions) performs at *44.7* % BLEU / *49.1* % TER with the same parallel training data and LM

Extensions to Hierarchical Machine Translation

Goals:

- ***Significantly improved translation quality*** **within large-scale Arabic→English system**
- ***Decoding speedups*** **without loss in translation performance**

Evaluated techniques:

- **Shallow rules**
- **IBM-style reorderings**
- **Soft syntactic labels**
- **Lightly-supervised training**
- **Discriminative word lexicon**

Related Work

- [Iglesias et al., EACL 2009]
  Shallow rules for efficient hierarchical phrase-based decoding
- [Vilar et al., WMT 2010]
  IBM-style reorderings for HPBT (German→English)
- [Stein et al., AMTA 2010]
  Syntactic extensions to HPBT (Chinese→English)
- [Schwenk, IWSLT 2008]
  Lightly-supervised training for phrase-based system (French→English)
- [Mauser et al., EMNLP 2009]
  Discriminative word lexicon model in phrase-based system

Shallow Rules

Idea:

- **Modification of the grammar to constrain the search space**
- ***Restriction of the depth of the hierarchical recursion to one***
- **No modifications to the decoder necessary**

Method:

- **Generic non-terminal X replaced by two distinct non-terminals XH and XP**
- **On all right-hand sides of hierarchical rules: XP**
- **Left-hand sides of lexical rules: XP**
- **Left-hand sides of hierarchical rules: XH**
- **Gaps within hierarchical phrases can thus only be filled with purely lexicalized phrases**

Shallow Rules: Initial and Glue Rule

- ***Initial rule*** **has to be substituted with two rules**

$$S \rightarrow \langle XP^{\sim 0}, XP^{\sim 0} \rangle$$
$$S \rightarrow \langle XH^{\sim 0}, XH^{\sim 0} \rangle$$

- ***Glue rule*** **has to be substituted with two rules**

$$S \rightarrow \langle S^{\sim 0} XP^{\sim 1}, S^{\sim 0} XP^{\sim 1} \rangle$$
$$S \rightarrow \langle S^{\sim 0} XH^{\sim 1}, S^{\sim 0} XH^{\sim 1} \rangle$$

IBM-Style Reorderings

Idea:

- **Include additional reorderings on top of the hierarchically motivated ones**

Method:

- ***Phrase-based IBM-style reorderings with a window length of 1***
- **Grammar-based implementation (replacement of initial and glue rule), with minimal modifications to the decoder**
- **Computation of distance-based jump cost**

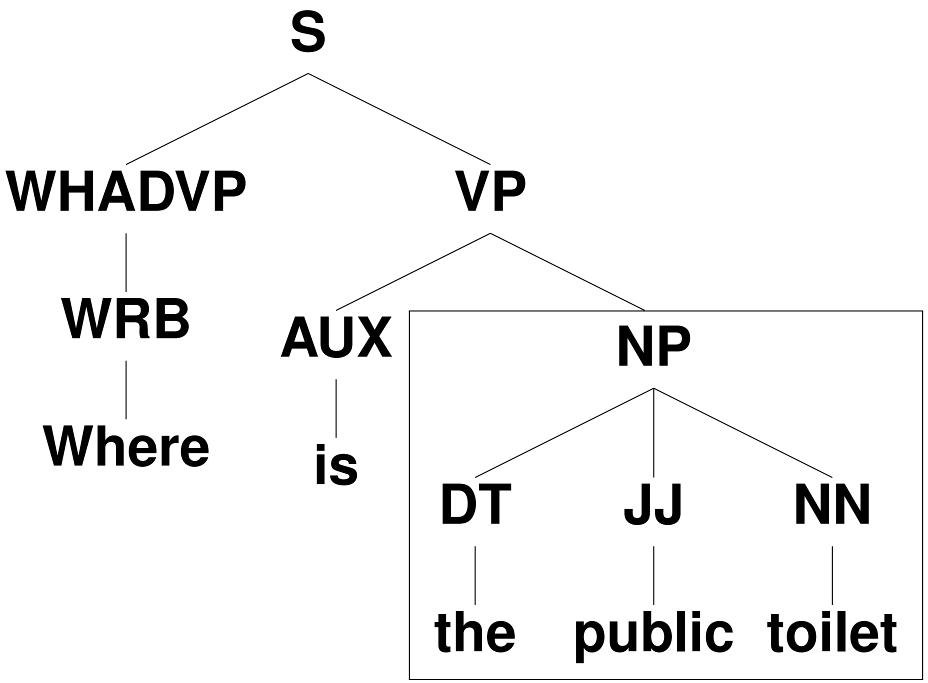
IBM-Style Reorderings: Initial and Glue Rule

$$
\begin{aligned}
S &\rightarrow \langle M^{\sim 0}, M^{\sim 0} \rangle \\
S &\rightarrow \langle M^{\sim 0} S^{\sim 1}, M^{\sim 0} S^{\sim 1} \rangle \\
S &\rightarrow \langle B^{\sim 0} M^{\sim 1}, M^{\sim 1} B^{\sim 0} \rangle \\
M &\rightarrow \langle X^{\sim 0}, X^{\sim 0} \rangle \\
M &\rightarrow \langle M^{\sim 0} X^{\sim 1}, M^{\sim 0} X^{\sim 1} \rangle \\
B &\rightarrow \langle X^{\sim 0}, X^{\sim 0} \rangle \\
B &\rightarrow \langle B^{\sim 0} X^{\sim 1}, B^{\sim 0} X^{\sim 1} \rangle
\end{aligned}
$$

- **M non-terminal represents a block that will be translated in a monotonic way**
- **B is a "back jump"**
- **Keep them separate for more flexibility (e.g. restriction of jump width)**

Soft Syntactic Labels: Principle

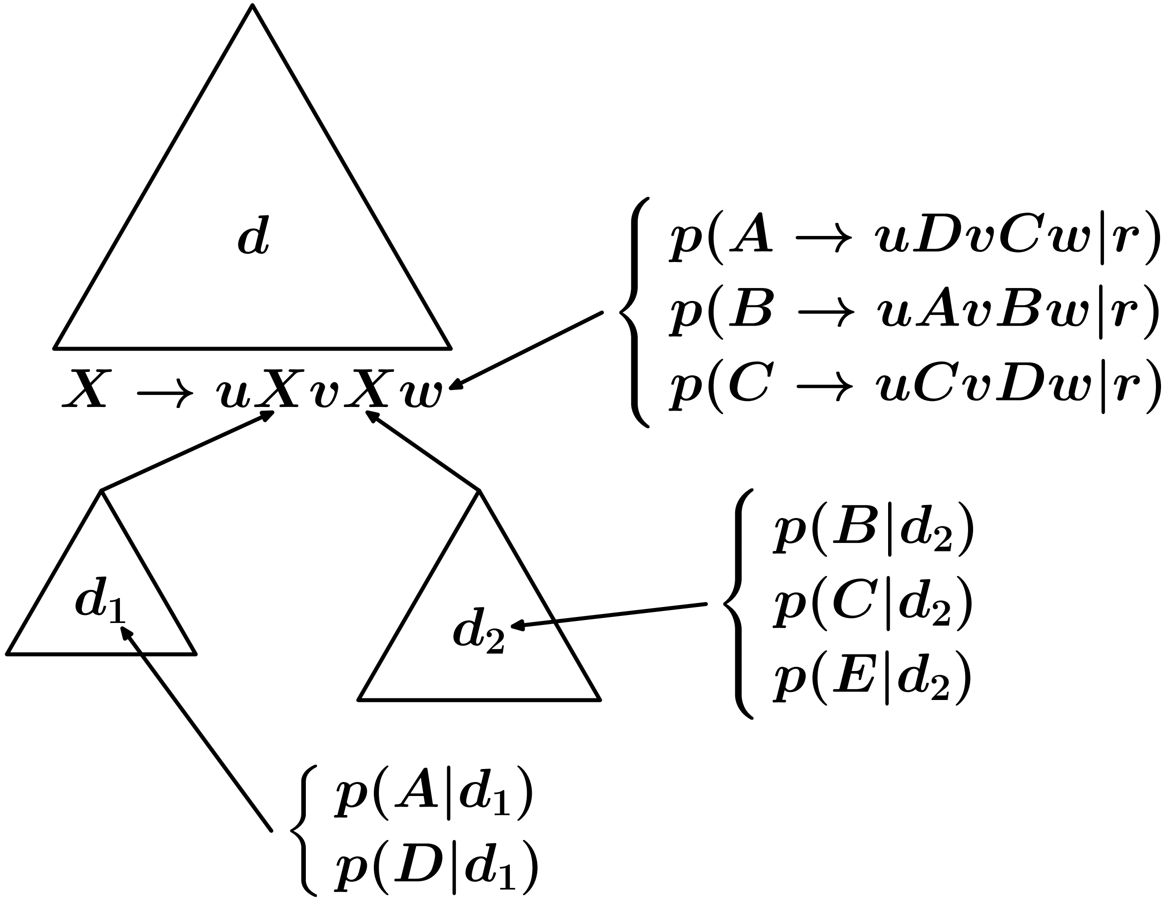
- Use *labels from syntactic parse trees* to replace the generic non-terminals in the translation process
- Target side of the training data is parsed (here: Berkeley Parser [Petrov et al. 2006])
- Resulting syntax trees are used in the rule extraction process



Soft Syntactic Labels: Model

- **Computation of two additional models for the log-linear combination**
  1. ***Tree well-formedness probability model*** p_{syntax} **for the parse tree construced by the decoder**
  2. ***Penalty for non-matching non-terminals***
- **Same phrase pairs, but syntax is stored as additional information in the rules**
- **Before: set of non-terminals** $\mathcal{NT} = \{S, X\}$
- **Now extended by a set of non-terminals in the additional model** $\mathcal{H} = \{NP, PP, NN, DT\ \ldots\}$

Soft Syntactic Labels: Decoding



- **$p(h_0|d_1)$ is a computed distribution over all labels $h_0 \in \mathcal{H}$ for sub-derivation d_1**
- **$p(\mathrm{h}|r)$ is the distribution computed in the rule extraction for rule r**

Lightly-Supervised Training

Idea:

- *Automatically translate monolingual source language corpora*
- **Create word alignments on resulting bitexts**
- *Use as unsupervised parallel training data*

Method:

- **Cross-system and cross-paradigm variant of lightly-supervised training**
  - Automatic translations of parts of the Arabic LDC Gigaword corpus
  - Created with a standard phrase-based system and kindly provided by Holger Schwenk, LIUM, Le Mans
  - Selection of *4.7M sentence pairs*
  - Used as additional training material for RWTH's HPBT system
- **Lexical phrases extracted from unsupervised data, hierarchical phrases from more reliable human-generated parallel data only**
- **Number of non-hierarchical phrases increased by roughly 30%**

Discriminative Word Lexicon (DWL)

Discriminative, log-linear lexicon model: $p(e|f_1^J)$

- ***Predict the words contained in the translation from the words given in the source sentence***
- **2-class classification problem: target word included / not included in translation**
- **Features: words in the source sentence**
- **Captures context beyond phrase boundaries and n-gram language model history**

Training:

- **Improved RProp+ [Igel & Hüsken 2003], L2-regularization**
- **Easy to parallelize: one target word per core**
- **But many parameters: weights for all source word / target word combinations**
- **Full model trained, threshold pruning applied afterwards to discard features with low values (separate for each class)**

DWL Training NIST Arabic→English

- DWL model trained on a high-quality subset of 0.3M sentence pairs
- RProp+: 100 iterations per target word
- Pruned with threshold 0.1
- On average 80 features per target word (unpruned: 122 592)

Experimental Results NIST Arabic→English

	Arabic → English (MT08)							
	deep				shallow			
	BLEU [%]		TER [%]		BLEU [%]		TER [%]	
HPBT Baseline	44.3	±1.1	50.0	±0.9	44.4	±1.1	49.4	±0.9
+ Unsup	45.0	+0.7	49.4	-0.6	45.2	+0.8	49.2	-0.2
+ Unsup + DWL	*45.7*	+1.4	*48.7*	-1.3	*45.8*	+1.4	*48.2*	-1.2
+ Unsup + Syntactic Labels	45.2	+0.9	49.3	-0.7	45.0	+0.6	49.0	-0.4
+ Unsup + Reorderings	45.3	+1.0	49.1	-0.9	45.3	+0.9	48.9	-0.5
+ Unsup + DWL + Syntactic Labels	*46.0*	+1.7	*48.2*	-1.8	*45.8*	+1.4	*48.3*	-1.1
+ Unsup + DWL + Reorderings	*45.7*	+1.4	*48.7*	-1.3	*45.9*	+1.5	*48.2*	-1.2

- The 95% confidence interval is given for the baseline systems
- Highlighted results are significantly better than the baseline

Summary

- Significant improvements in Arabic $\rightarrow$ English HPBT due to
  - ***lightly-supervised training***
  - ***a discriminative word lexicon***
- **Decoding speedups (factor 5-10) without loss in translation quality with *shallow rules***
- ***Soft syntactic labels* and additional IBM-style *reorderings* have little to no impact**

Thank you for your attention

Matthias Huck

huck@cs.rwth-aachen.de

http://www-i6.informatik.rwth-aachen.de/