A Class-Based Agreement Model for Generating Accurately Inflected Translations

ACL 2012 // Jeju

Spence Green

Stanford University

John DeNero

Google

Local Agreement Error

Input: *The car goes quickly.*

Reference:

(1) a. السيارة تذهب بسرعة
the-car$_{+F}$ go$_{+F}$ with-speed

Local Agreement Error

Input: *The car goes quickly.*

Reference:

(1) a. السيارة تذهب بسرعة
the-car$_{+F}$ go$_{+F}$ with-speed

Google Translate:

(2) a. السيارة يذهب بسرعة
the-car$_{+F}$ go$_{+M}$ with-speed

Long-distance Agreement Error

Input: *The one who is speaking is my wife.*

Reference:

(3) a. celle qui parle , c'est ma femme
one$_{+F}$ who speak , is my wife$_{+F}$

Long-distance Agreement Error

Input: *The one who is speaking is my wife.*

Reference:

(3) a. celle qui parle , c'est ma femme
one$_{+F}$ who speak , is my wife$_{+F}$

Google Translate:

(4) a. celui qui parle est ma femme
one$_{+M}$ who speak is my spouse$_{+F}$

Agreement Errors: Really Annoying

Ref John runs to his house.

MT John run to her house.

Agreement Errors in Phrase-Based MT

Agreement relations cross phrase boundaries

بسرعة يذهب	السيارة

Agreement Errors in Phrase-Based MT

Agreement relations cross phrase boundaries

السيارة	بسرعة يذهب

Language model should help?

- Sparser *n*-gram counts
- LM may back off more often

Possible Solutions

Morphological generation e.g. [Minkov et al. 2007]

- Useful when correct translations aren't in phrase table

Possible Solutions

Morphological generation e.g. [Minkov et al. 2007]

- Useful when correct translations aren't in phrase table

Our work: model agreement with a new feature

- Large phrase tables already contain many word forms

Key Idea: Morphological Word Classes

السيارة ‘car’

$$\begin{bmatrix} \text{CAT} & \textit{noun} \\ \text{AGR} & \begin{bmatrix} \text{GEN} & \textit{fem} \\ \text{NUM} & \text{sg} \end{bmatrix} \end{bmatrix}$$

تذهب ‘to go’

$$\begin{bmatrix} \text{CAT} & \textit{verb} \\ \text{AGR} & \begin{bmatrix} \text{GEN} & \textit{fem} \\ \text{NUM} & \text{sg} \\ \text{PER} & 3 \end{bmatrix} \end{bmatrix}$$

Key Idea: Morphological Word Classes

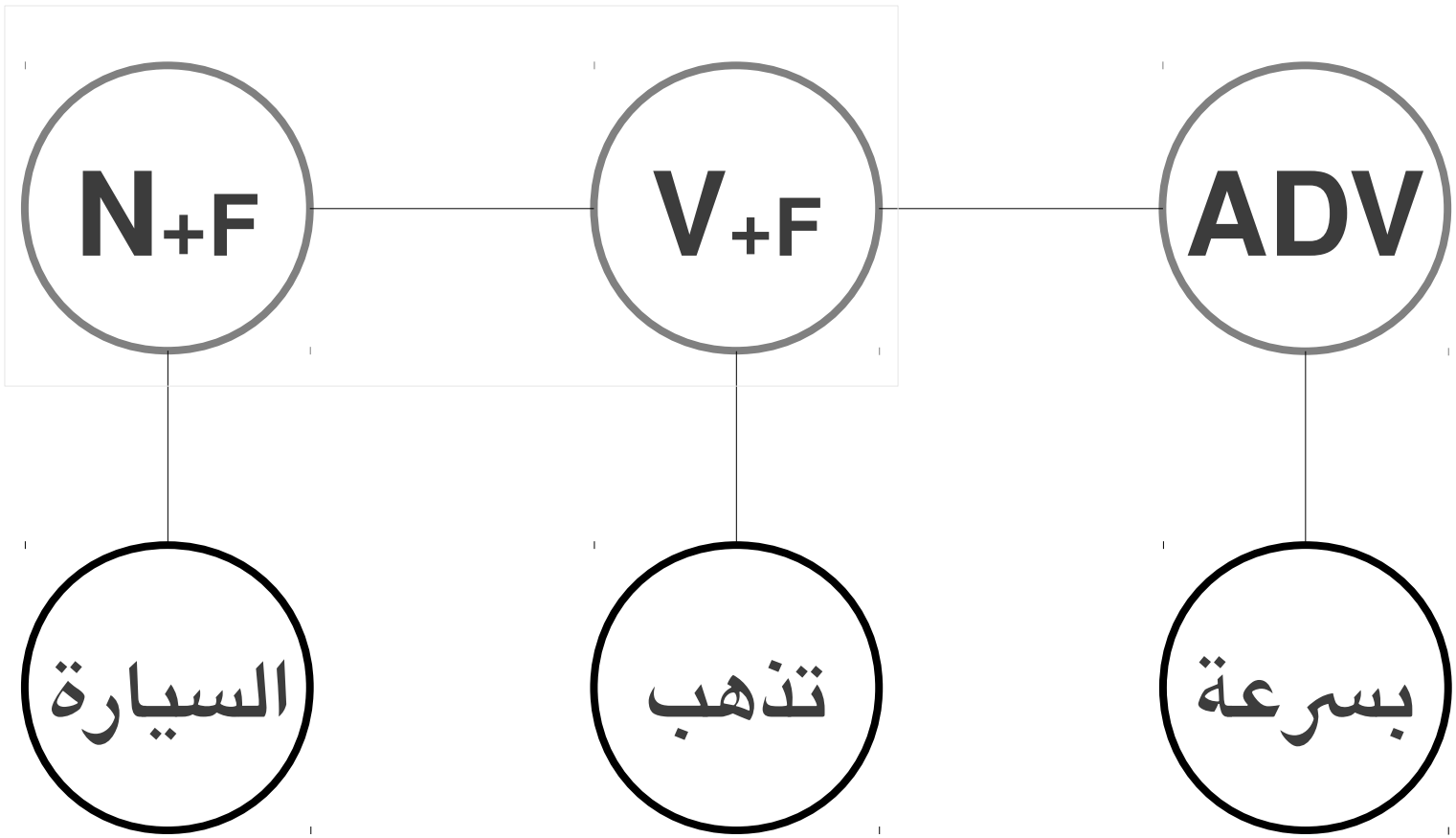
السيارة ‘car’	**noun+fem+sg**
تذهب ‘to go’	**verb+fem+sg+3**

Key Idea: Morphological Word Classes

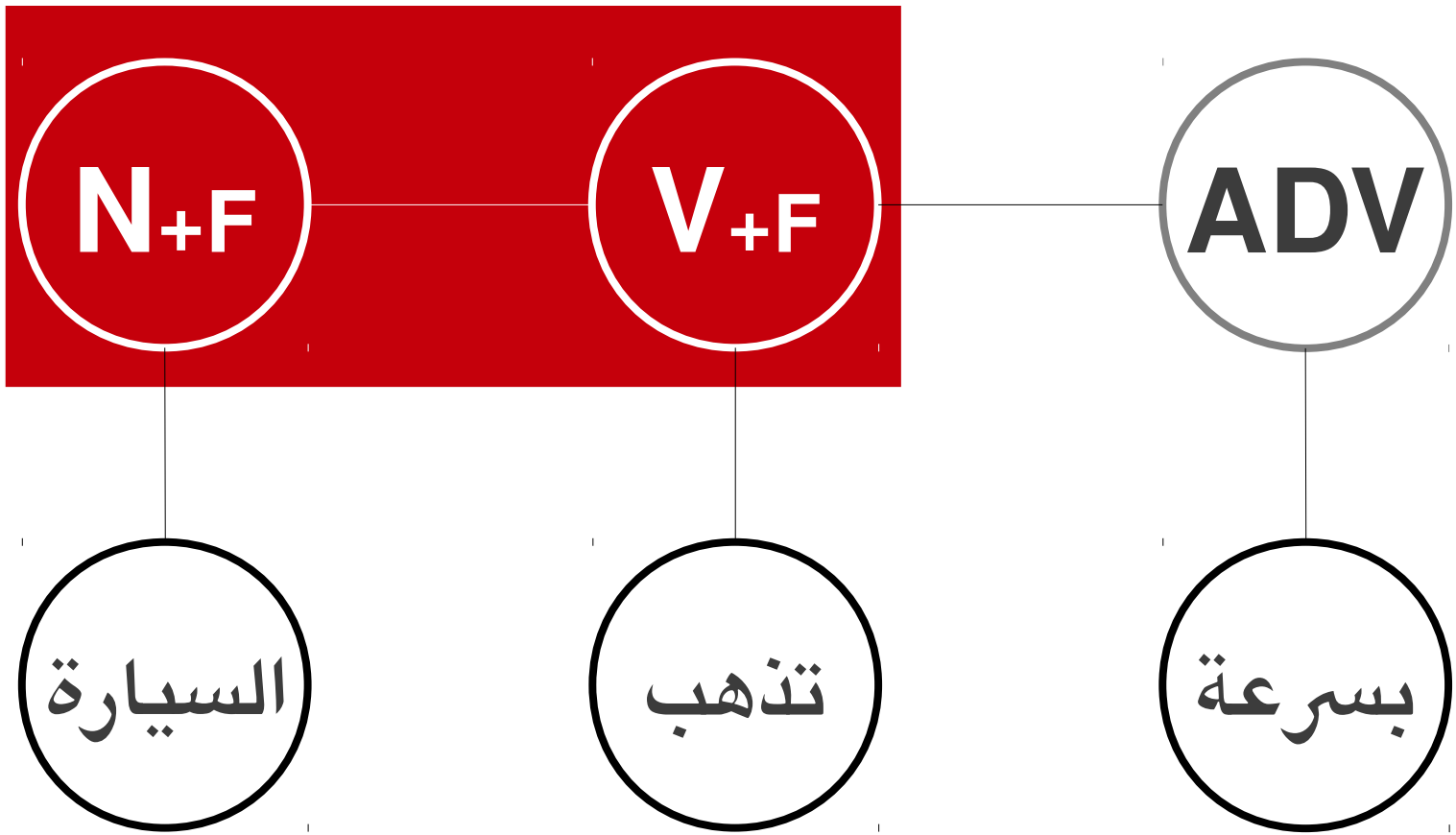
السيارة ‘car’	**noun+fem+sg**
تذهب ‘to go’	**verb+fem+sg+3**

Linearized feature structure is equally expressive, assuming a fixed order

A Class-based Agreement Model



A Class-based Agreement Model



1. Model Formulation (for Arabic)
2. MT Decoder Integration
3. English-Arabic Evaluation

1. Model Formulation (for Arabic)

2. MT Decoder Integration

3. English-Arabic Evaluation

Agreement Model Formulation

Implemented as a decoder feature

Agreement Model Formulation

Implemented as a decoder feature

when **each** hypothesis $h \in \mathcal{H}$ is extended:

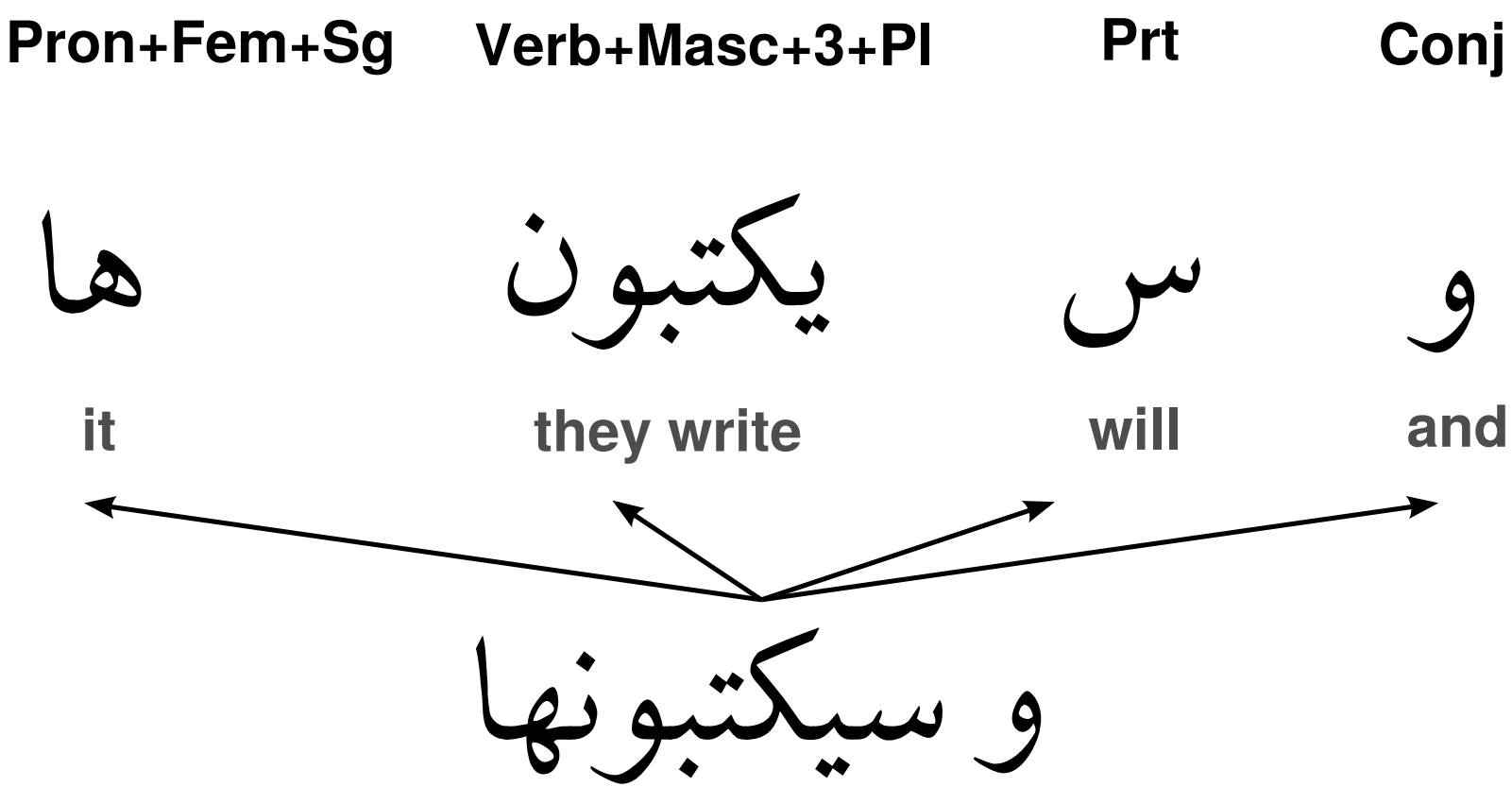
$\hat{s} = \text{segment}(h)$

$\tau = \text{tag}(\hat{s})$

$q(h) = \text{score}(\tau)$

return $q(h)$

Step 1: Segmentation



Step 1: Segmentation

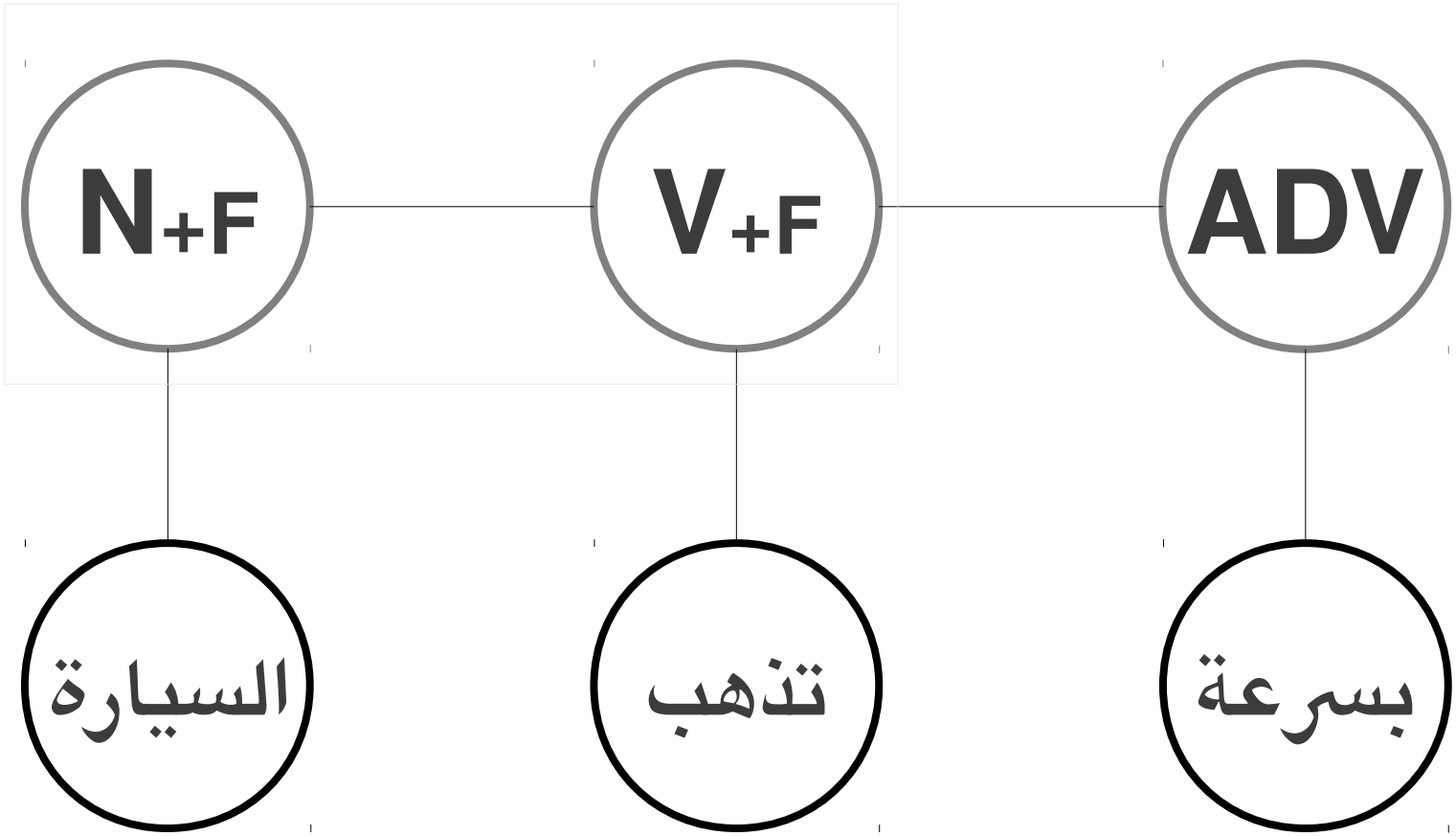
Character-level CRF: $p(\hat{s}|words)$

Features: Centered 5-character window

Label set:

- **I** inside segment
- **O** outside segment (whitespace)
- **B** beginning of segment
- **F** do not segment (punctuation, digits, ASCII)

Step 2: Tagging



Step 2: Tagging

Token-level CRF: $p(\tau|\hat{s})$

Features: Current and previous words, affixes, etc.

Label set: morphological classes (89 for Arabic)

- Gender, number, person, definiteness

Step 2: Tagging

Token-level CRF: $p(\tau|\hat{s})$

Features: Current and previous words, affixes, etc.

Label set: morphological classes (89 for Arabic)

- Gender, number, person, definiteness

What about incomplete hypotheses?

Step 3: Scoring

Problem: Discriminative model score $p(\tau|\hat{s})$ not comparable across hypotheses

- MST parser score: **works?** [Galley and Manning 2009]
- CRF score: **fail** [this paper]

Step 3: Scoring

Problem: Discriminative model score $p(\tau|\hat{s})$ not comparable across hypotheses

- MST parser score: **works?** [Galley and Manning 2009]
- CRF score: **fail** [this paper]

Solution: Generative scoring of class sequences

Step 3: Scoring

Simple bigram LM trained on gold class sequences

$$\tau^* = \arg\max_{\tau} p(\tau|\hat{s})$$

$$q(h) = p(\tau^*) = \prod_i p(\tau_i^*|\tau_{i-1}^*)$$

Step 3: Scoring

Simple bigram LM trained on gold class sequences

$$\tau^* = \arg\max_{\tau} p(\tau|\hat{s})$$

$$q(h) = p(\tau^*) = \prod_i p(\tau_i^*|\tau_{i-1}^*)$$

Order of scoring model dependent on MT decoder design

1. Model Formulation (for Arabic)

2. MT Decoder Integration

3. English-Arabic Evaluation

MT Decoder Integration

Tagger CRF

1. Remove next-word features
2. Only tag boundary for goal hypotheses

MT Decoder Integration

Tagger CRF

1. Remove next-word features
2. Only tag boundary for goal hypotheses

Hypothesis state: last segment + class

LM history: السيارة تذهب

Agreement history: تذهب / verb+fem+sg+3

1. Model Formulation (for Arabic)
2. MT Decoder Integration
3. English-Arabic Evaluation

Component Models (Arabic Only)

	Full (%)	Incremental (%)
Segmenter	98.6	–
Tagger	96.3	96.2

Data: Penn Arabic Treebank [Maamouri et al. 2004]

Setup: Dev set, standard split [Rambow et al. 2005]

Translation Quality

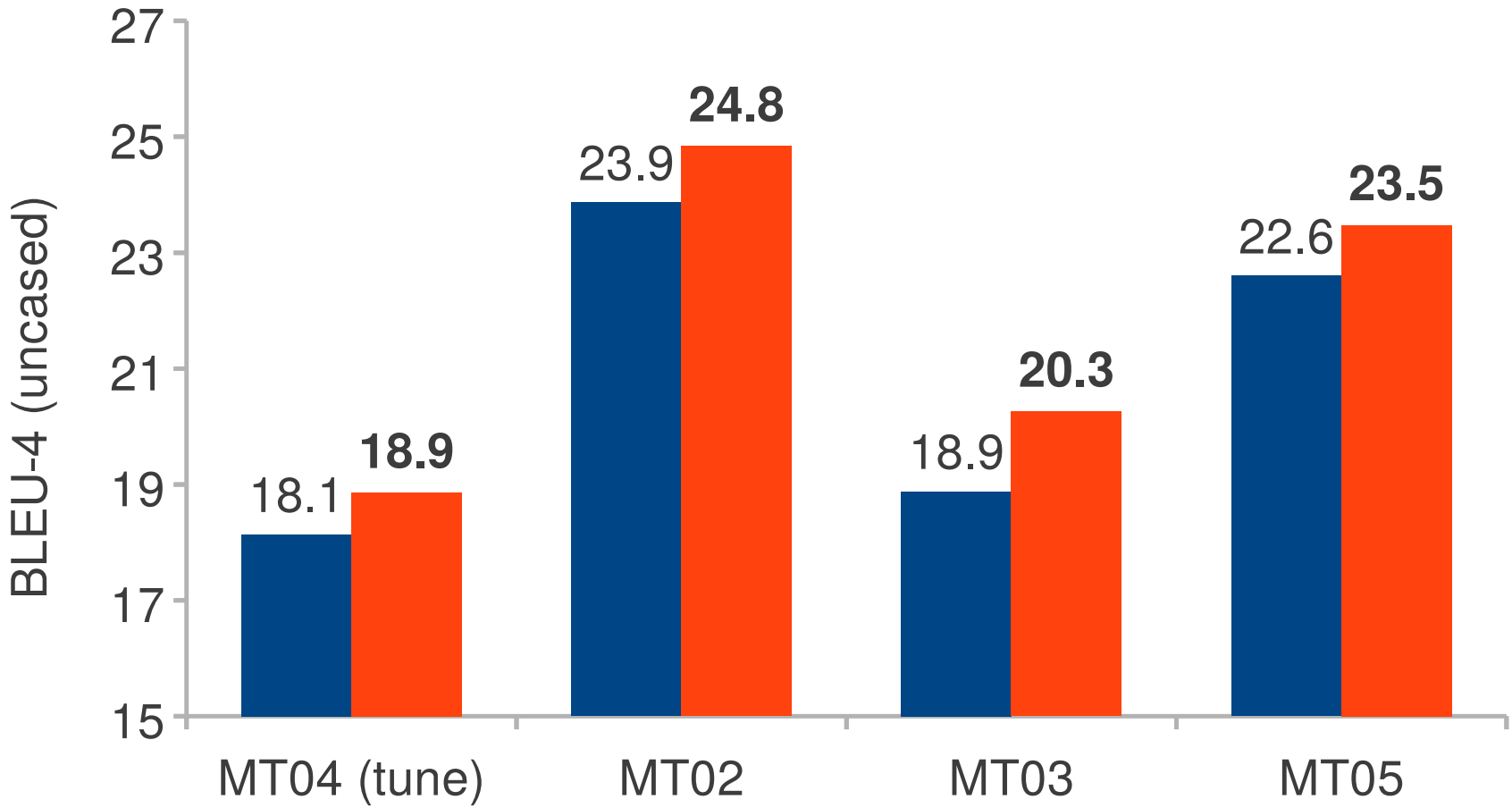
Phrase-based decoder [Och and Ney 2004]

- Phrase frequency, lexicalized re-ordering model, etc.

Bitext: 502M English-Arabic tokens

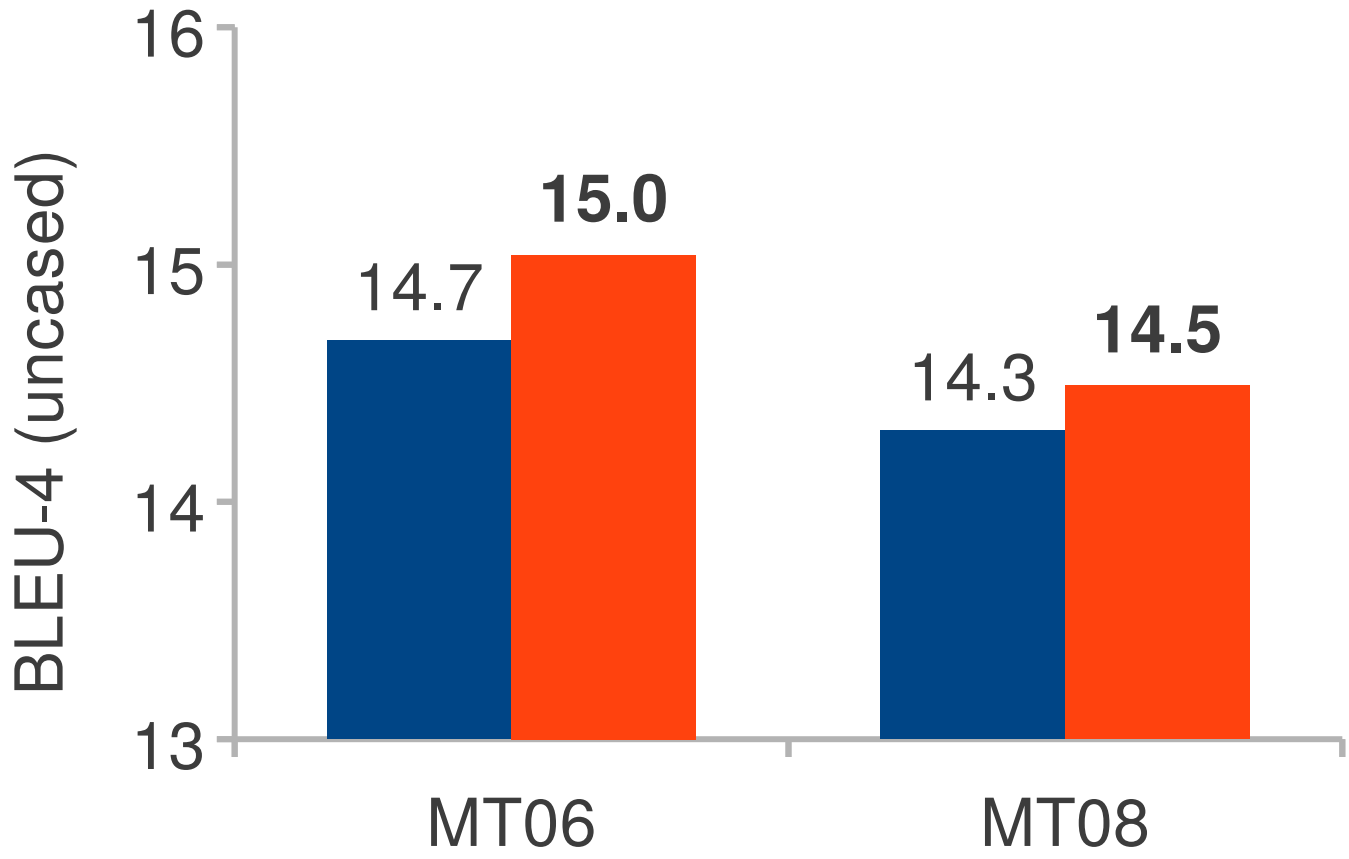
LM: 4-gram from 600M Arabic tokens

Translation Quality: NIST Newswire



Average gain: +1.04 BLEU (significant at $p \leq 0.01$)

Translation Quality: NIST Mixed Genre



Average gain: +0.29 BLEU (significant at $p \leq 0.02$)

Human Evaluation

MT05 output: 74.3% of hypotheses differed from baseline

Sampled 100 sentence pairs

Manually counted agreement errors

Human Evaluation

MT05 output: 74.3% of hypotheses differed from baseline

Sampled 100 sentence pairs

Manually counted agreement errors

Result: 15.4% error reduction, $p \leq 0.01$ (78 vs. 66)

Analysis: Phrase Table Coverage

Hypothesis: Inflected forms in phrase table, but unused

Analysis: Phrase Table Coverage

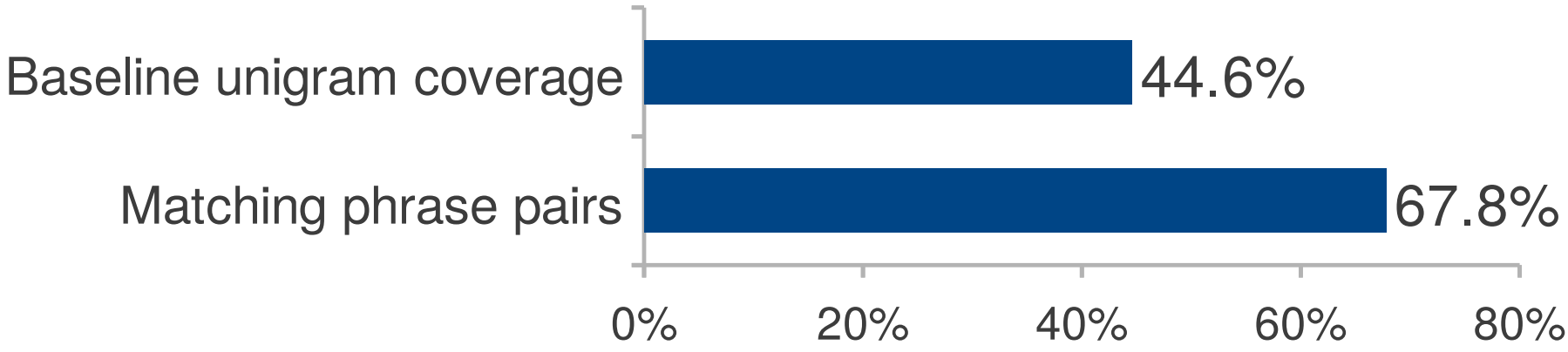
Hypothesis: Inflected forms in phrase table, but unused

Analysis: Measure MT05 reference unigram coverage

Analysis: Phrase Table Coverage

Hypothesis: Inflected forms in phrase table, but unused

Analysis: Measure MT05 reference unigram coverage



Conclusion: Implementation is Easy

You need:

1. CRF package
2. Know-how for implementing decoder features
3. Morphologically annotated corpus

Conclusion: Contributions

Translation quality improvement in a large-scale system

Conclusion: Contributions

Translation quality improvement in a large-scale system

Classes and segmentation predicted during decoding

- Modeling flexibility

Conclusion: Contributions

Translation quality improvement in a large-scale system

Classes and segmentation predicted during decoding

- Modeling flexibility

Foundation for structured language models

- Future work: long-distance relations

Segmenter: `nlp.stanford.edu/software/`

thanks.

ولكم جزيل الشكر

References

Galley, M. and C. D. Manning (2009). “Quadratic-time dependency parsing for machine translation”. In: *ACL-IJCNLP*.
Maamouri, M. et al. (2004). “The Penn Arabic Treebank: Building a large-scale annotated Arabic corpus”. In: *NEMLAR*.
Minkov, E., K. Toutanova, and H. Suzuki (2007). “Generating Complex Morphology for Machine Translation”. In: *ACL*.
Och, F. J. and H. Ney (2004). “The alignment template approach to statistical machine translation”. In: *Computational Linguistics* 30.4, pp. 417–449.
Rambow, O. et al. (2005). *Parsing Arabic Dialects*. Tech. rep. Johns Hopkins University.