CLIR for Informal Content in Arabic Forum Posts

Mossaab Bagdouri, **Douglas W. Oard**   University of Maryland, College Park, MD, USA   mossaab@umd.edu
oard@umd.edu

Vittorio Castelli   IBM T.J. Watson Research Center, Yorktown Heights, NY, USA   vittorio@us.ibm.com
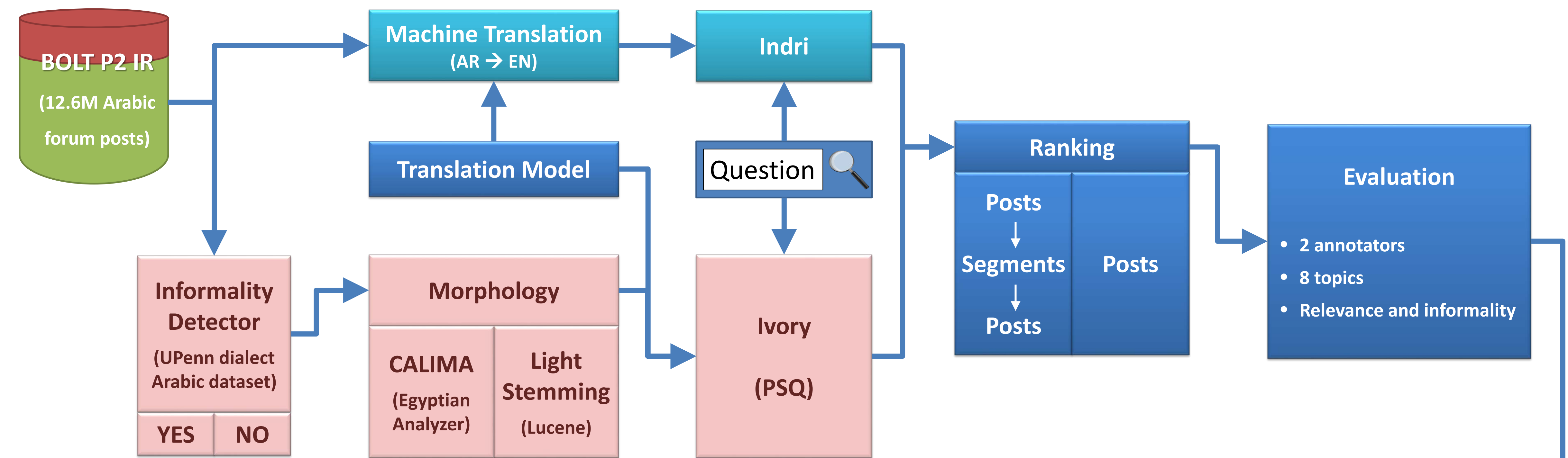
Introduction

Arabic Corpus — Formal / Informal

What do people say about tourism in Egypt after the revolution?

→ Results → **Current**

→ Goal → **Goal**

وقال إن ذلك سيؤدي إلي ارتفاع تكلفة الأنتاج للعصائر في مصر بنسبة في حدود 20% (...)

MT: He said that this would lead a rise in the cost of production for juices in Egypt at a rate of approximately 20 borders (...)

كلامها صحيح فعلا السياحة باى باى لسه **خطفيين** اتنين سياح في سيناء

MT: Its really true words tourism _any_ _any_ **Khatfin** just two tourists in Sinai

Relevant Informality

- Goal: optimize retrieval of informal content
- New measure: relevant and informal

Relevant

Informal

Conclusion & Future Work

New measure

New test collection

✓ Informality classification

✗ Probabilistic Structured Queries

✗ Dialect-tuned morphological analysis

? Pseudo-Relevance Feedback

? Dialect to MSA MT system

? Informal questions

? Query translation

Experiments

BOLT P2 IR (12.6M Arabic forum posts)

Machine Translation (AR → EN)

Indri

Translation Model

Question

Informality Detector (UPenn dialect Arabic dataset) — YES | NO

Morphology — CALIMA (Egyptian Analyzer) | Light Stemming (Lucene)

Ivory (PSQ)

Ranking: Posts → Segments → Posts
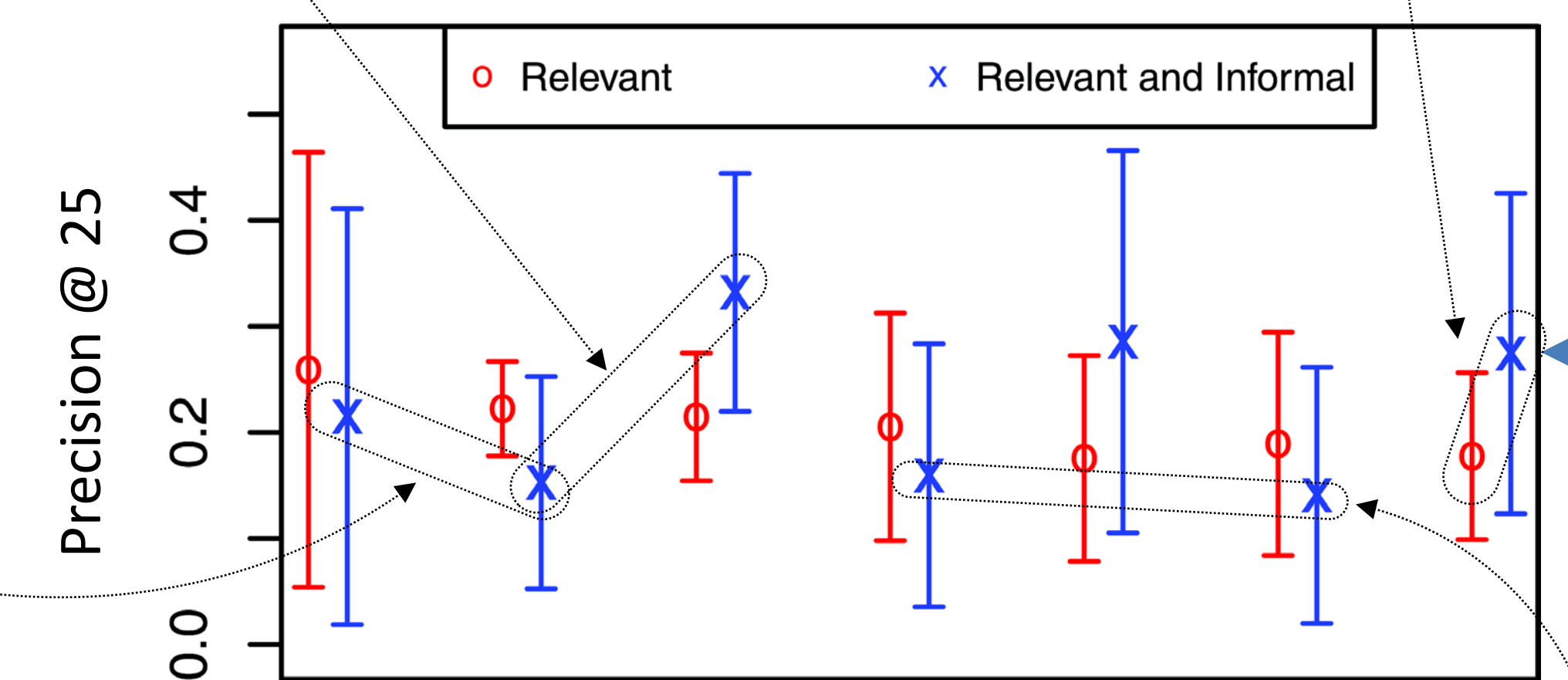
Evaluation
- 2 annotators
- 8 topics
- Relevance and informality

New measure rewards systems tuned to retrieve informality

Informality classifier improves precision significantly

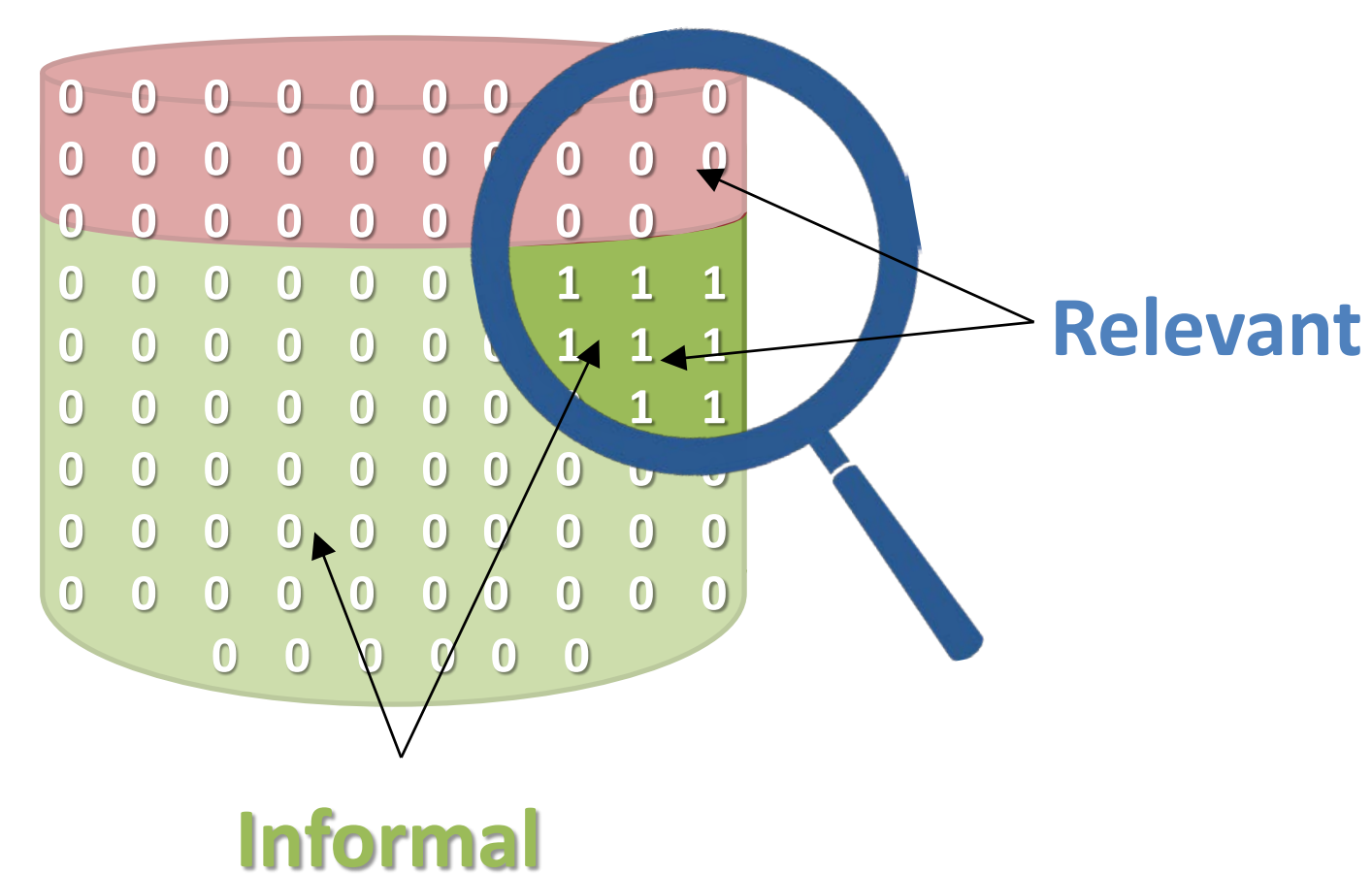No statistical difference between MT First and IR First

No difference between CALIMA and light stemming

Precision @ 25

○ Relevant   × Relevant and Informal

Informality detector	-		NO	YES	NO	YES	NO	YES
Morphology			CALIMA		CALIMA		Light Stemming	
Ranking	Posts → Segments → Posts				Posts			
CLIR	MT First		IR First (PSQ)					