TECHNISCHE
UNIVERSITÄT
WIEN
Vienna University of Technology

Effects of Language and Topic Size in Patent IR: An Empirical Study

Florina Piroi, Mihai Lupu, Allan Hanbury

Motivation

Evaluating IR engines in large-scale settings

- Lots of data
- Lots of topics
- Relevance judgments

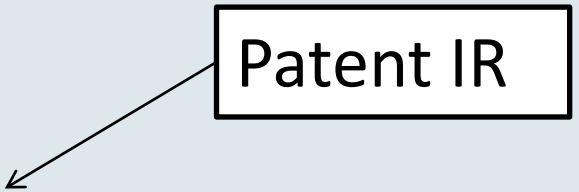
Motivation

Evaluating IR engines in IP, large-scale settings

Lots of topics

Relevance judgments

based on patent citations in search reports



Motivation

Evaluating IR engines in large-scale settings

Lots of topics ← Patent IR

Relevance judgments are few!

based on patent citations in search reports

Motivation

Evaluating IR engines in large-scale settings

Right topics

Relevance judgments are few!

based on patent citations in search reports

Motivation

How do retrieval results vary when we vary the set of topics?

Right topics

Patent IR

Relevance judgments are few!

based on patent citations in search reports

Motivation

How do retrieval results vary when we **vary** the set of topics?

Right topics

Patent IR

Relevance judgments are few!

based on patent citations in search reports

Document language
Document length

Document classification codes
Links to images
…

Experiment Setting

Choose patent-based collections

- CLEF-IP
- TREC-Chem

Experiment Setting

Choose patent-based collections

Document structure is the same

Prior Art Task

CLEF-IP

TREC-Che

Topic documents: complete patent document
Question: “Find documents that potentially invalidate the topic document”

Experiment Setting

English only

	CLEF-IP		TREC-Chem	
	2009	2010	2009	2010
Language	√	√		
Length (#words)	√	√	√	√

Experiment Setting

English only

	CLEF-IP		TREC-Chem	
	2009	2010	2009	2010
Language	√	√		
Length (#words)	√	√	√	√
#Topics	1000	1937	1000	1000
#runs	24	18	15	9

Experiment Setting

English only

	CLEF-IP		TREC-Chem	
	2009	2010	2009	2010
La				
Le (#			√	√
#Topics	1000	1937	1000	1000
#runs	24	18	15	9

MAP, NDCG
τ, ρ Correlations

Topic Document Size

- Generally, longer narratives do better
- IP Domain?

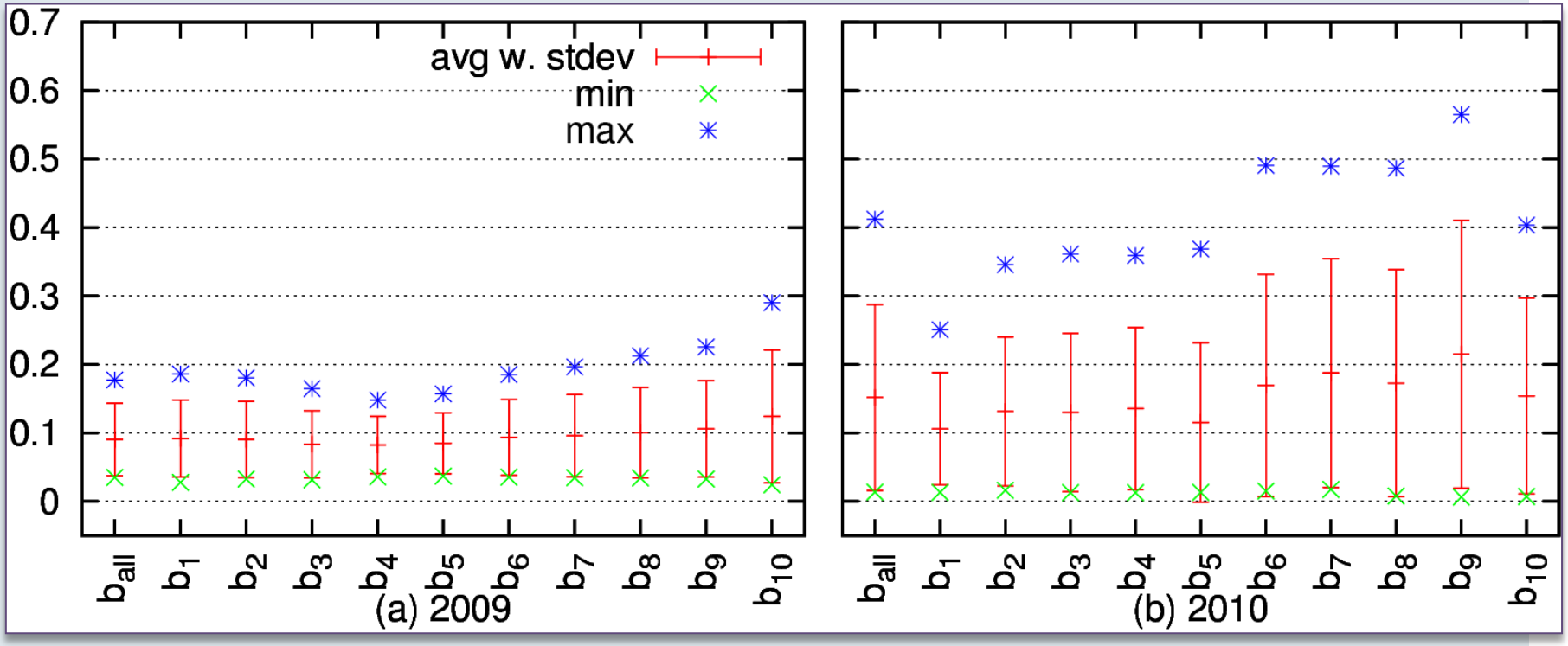
Topic Document Size

- Generally, longer narratives do better
- IP Domain?
- Split the topic set based on document length

Topic Document Size

MAP, TREC-Chem

- Generally, longer narratives do better



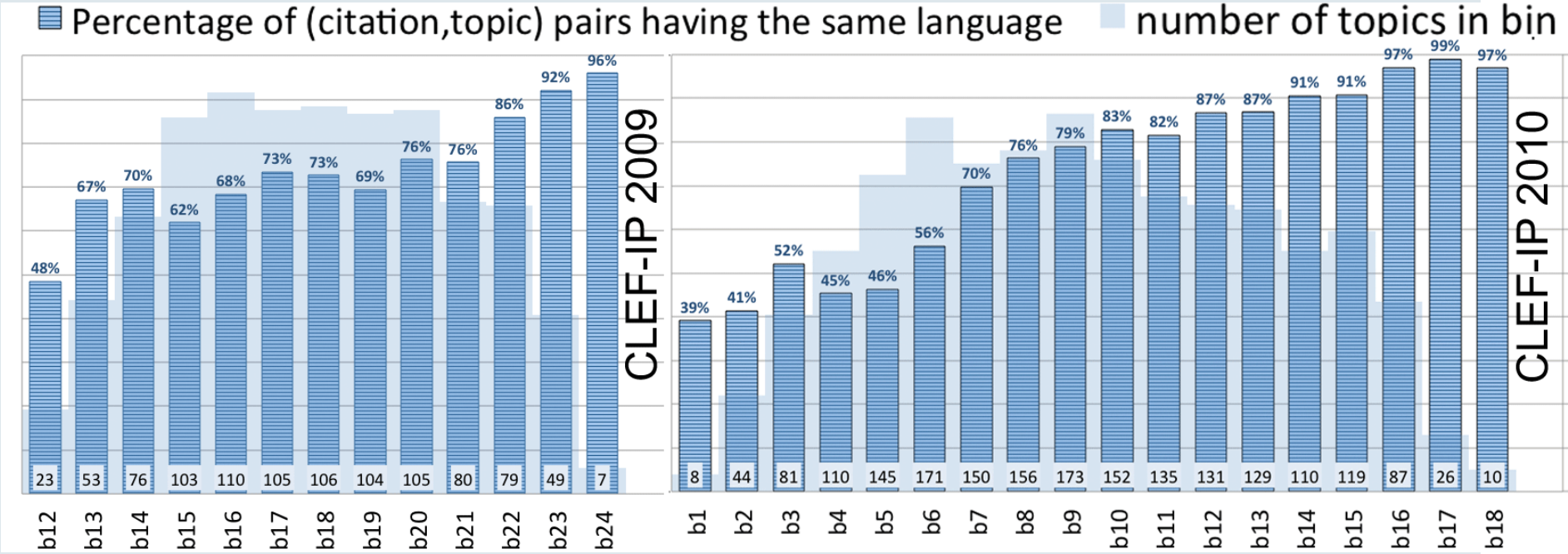
Topic Document Size

- Generally, longer narratives do better
- IP Domain?
- TREC-Chem – rather yes
- CLEF-IP – no proof found

Document Language

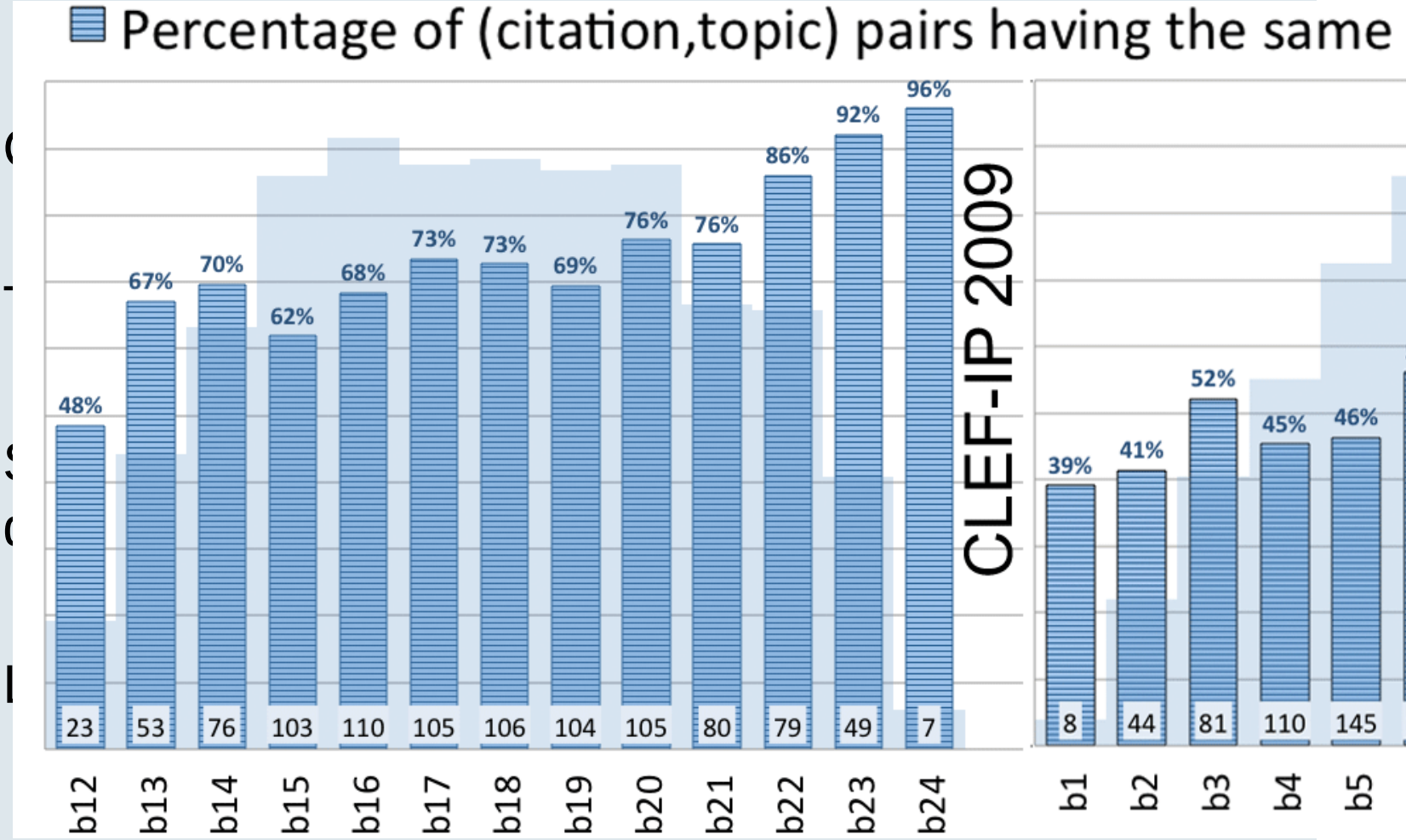
- CLEF-IP only
- Topic_language ≠ RelDoc_language
- Split the topic set by #runs that found all relevant documents → 24, 18 bins
- Look at ‹Topic_language, RelDoc_language› (each bin)

Document Language

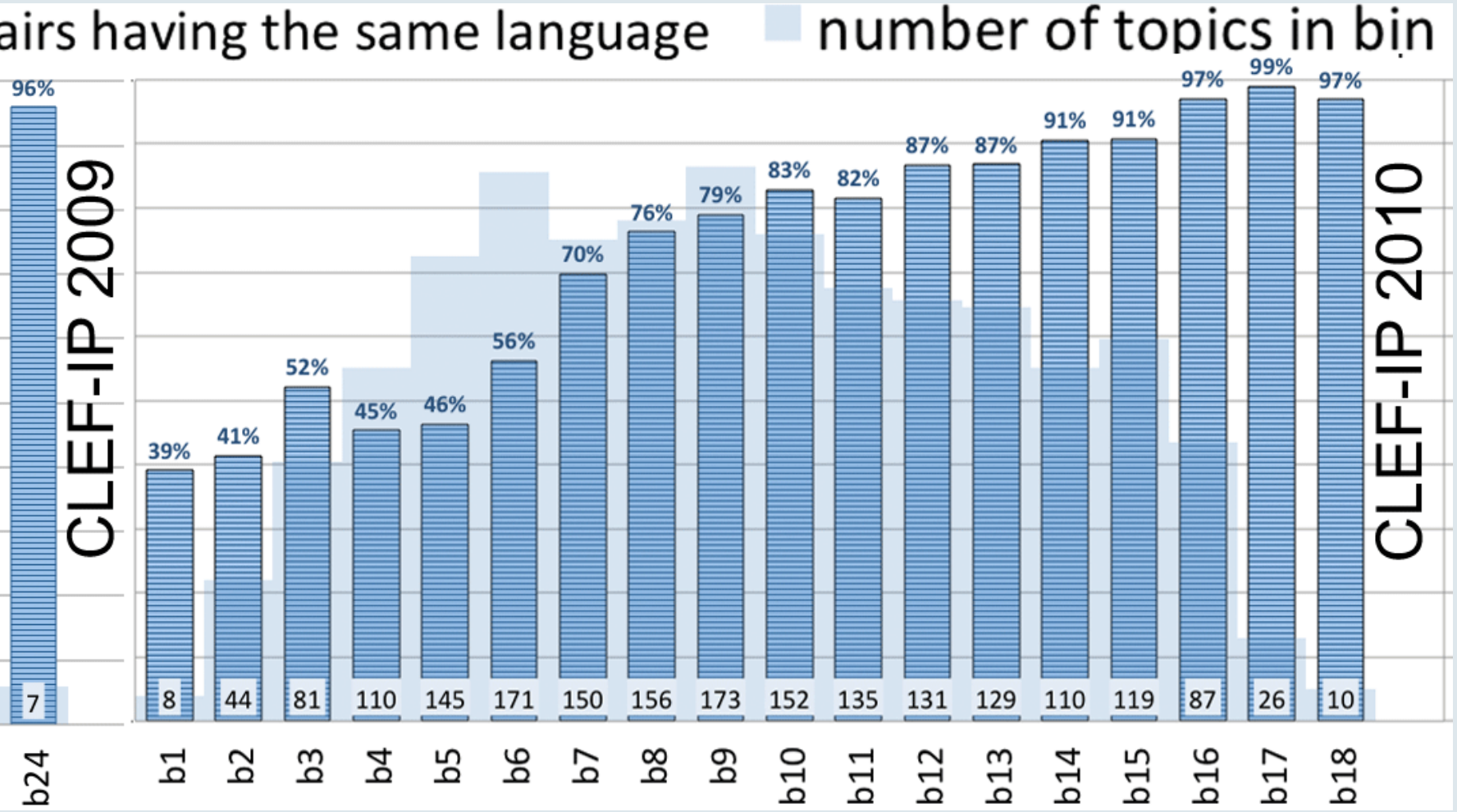


- Look at ‹Topic_language, RelDoc_language› (each bin)

Document Language



Document Language



Document Language

- It does matter! (we knew that)
- Top runs had better values for non English topics

Motivation

Experiment Settings

Results

Related Work

TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

Effects of Language and Topic Size in Patent IR: An Empirical Study

Florina Piroi, Mihai Lupu, Allan Hanbury

Thank You!