

# Cross-Language Retrieval

LBSC 796/INFM 718R

Douglas W. Oard

Session 12: April 27, 2011

# Agenda

- Questions
- Overview
- Cross-Language Search
- User Interaction

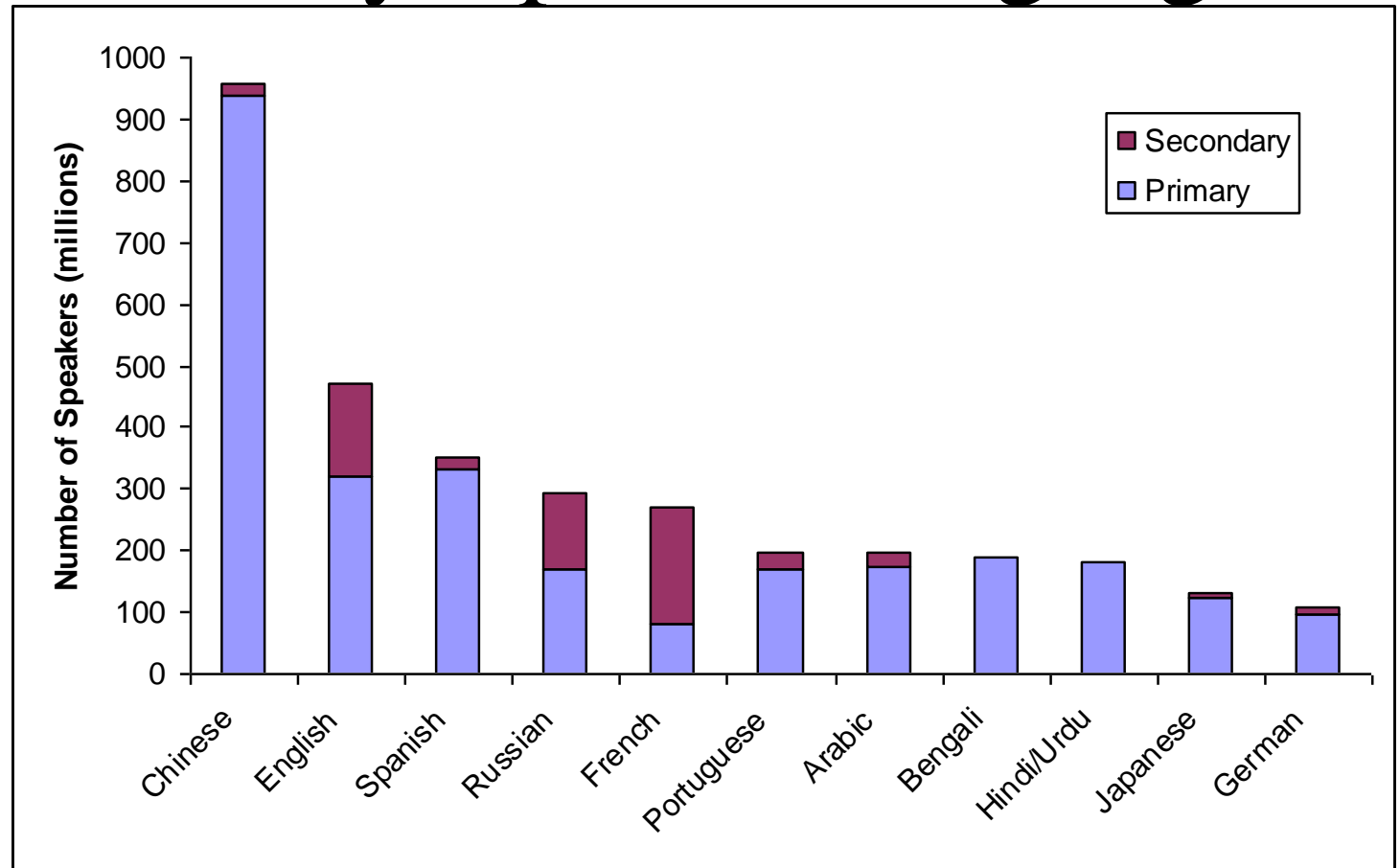
# User Needs Assessment

- Who are the potential users?
- What goals do we seek to support?
- What language skills must we accommodate?

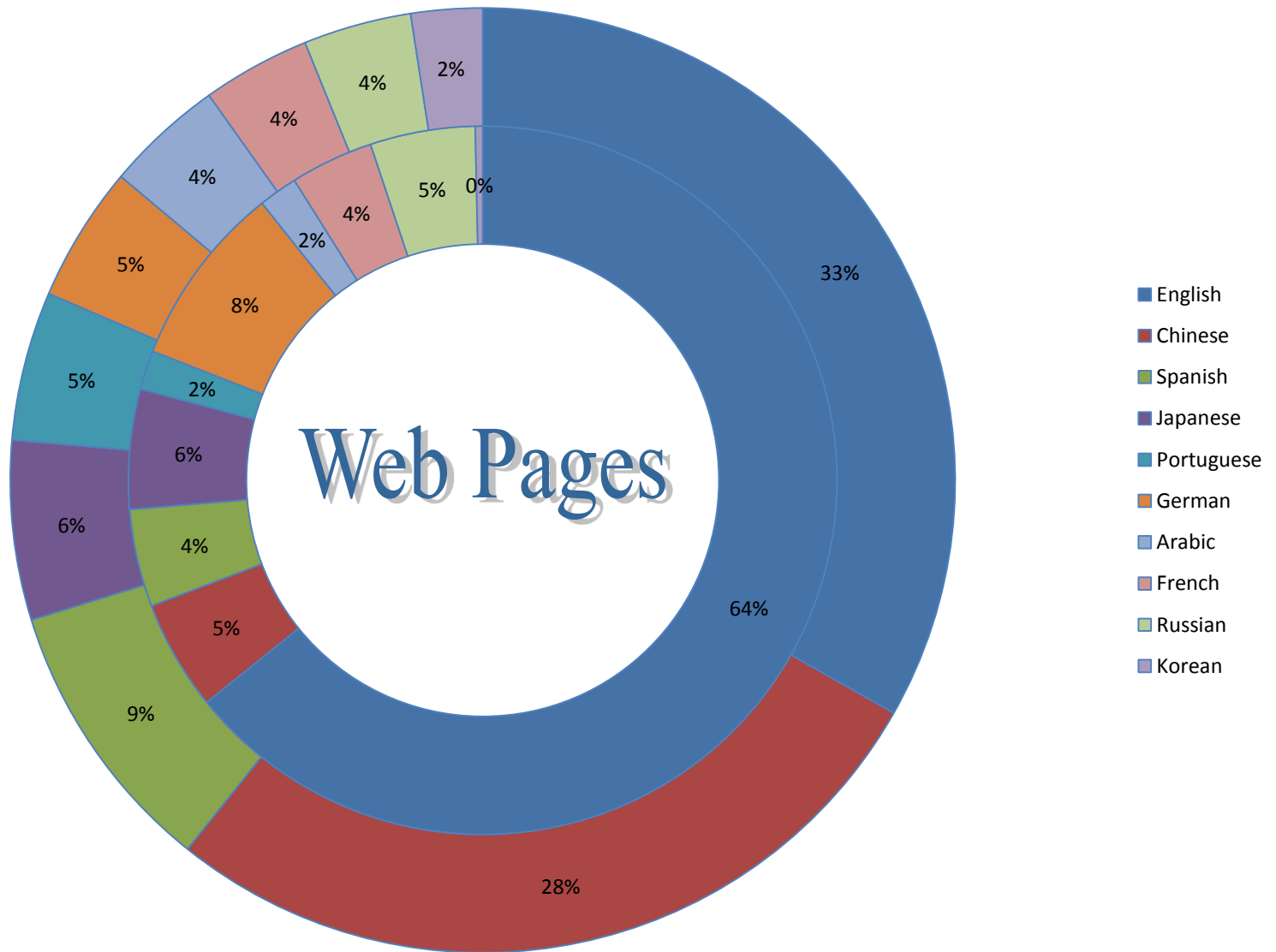
# Who needs Cross-Language Search?

- When users can read several languages
  - Eliminate multiple queries
  - Query in most fluent language
- Monolingual users can also benefit
  - If translations can be provided
  - If it suffices to know that a document exists
  - If text captions are used to search for images

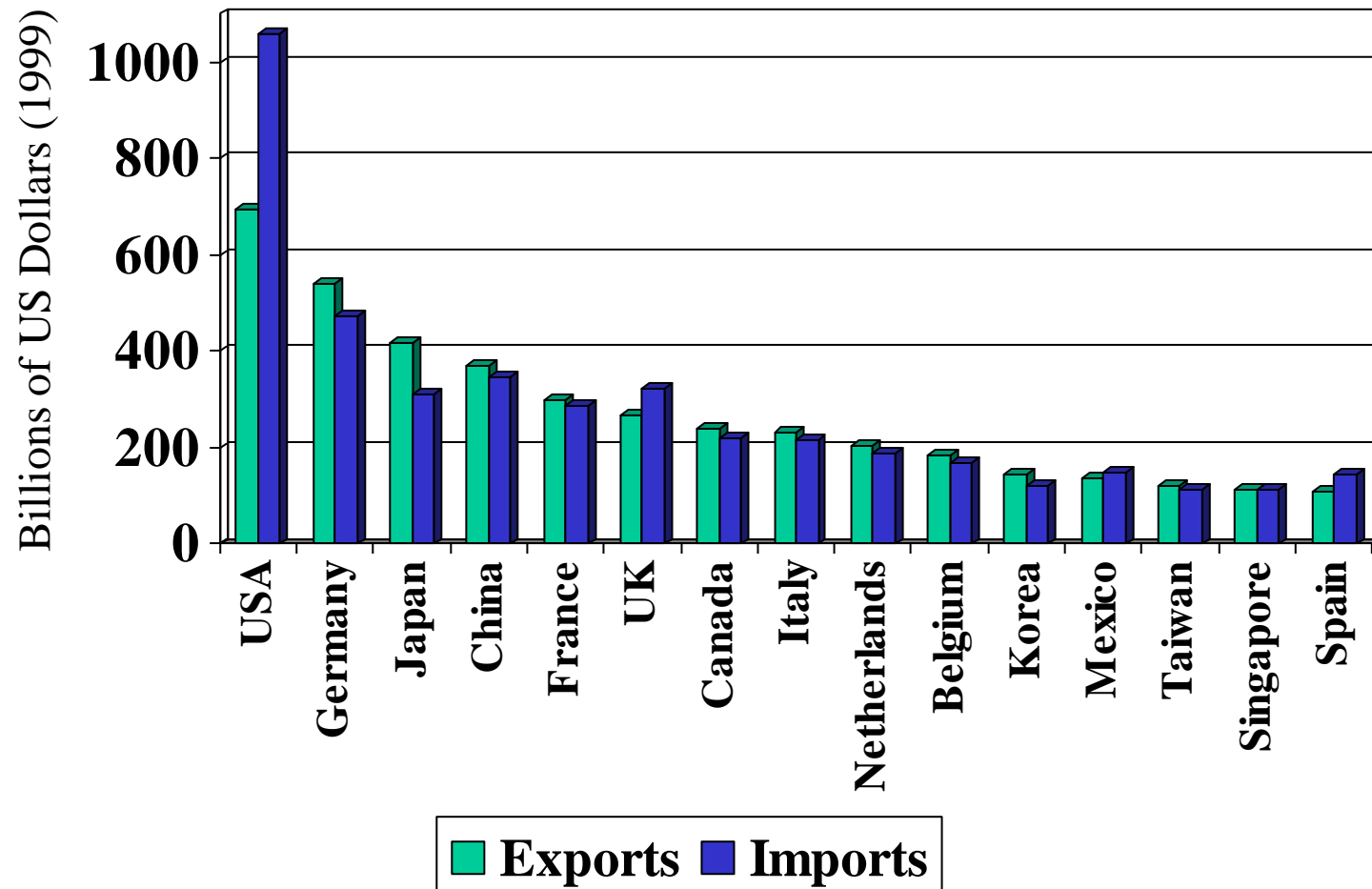
# Most Widely-Spoken Languages



# Global Internet Users



# Global Trade



Source: World Trade Organization 2000 Annual Report

# The Problem Space

- Retrospective search
  - Web search
  - Specialized services (medicine, law, patents)
  - Help desks
- Real-time filtering
  - Email spam
  - Web parental control
  - News personalization
- Real-time interaction
  - Instant messaging
  - Chat rooms
  - Teleconferences

## Key Capabilities

Map across languages

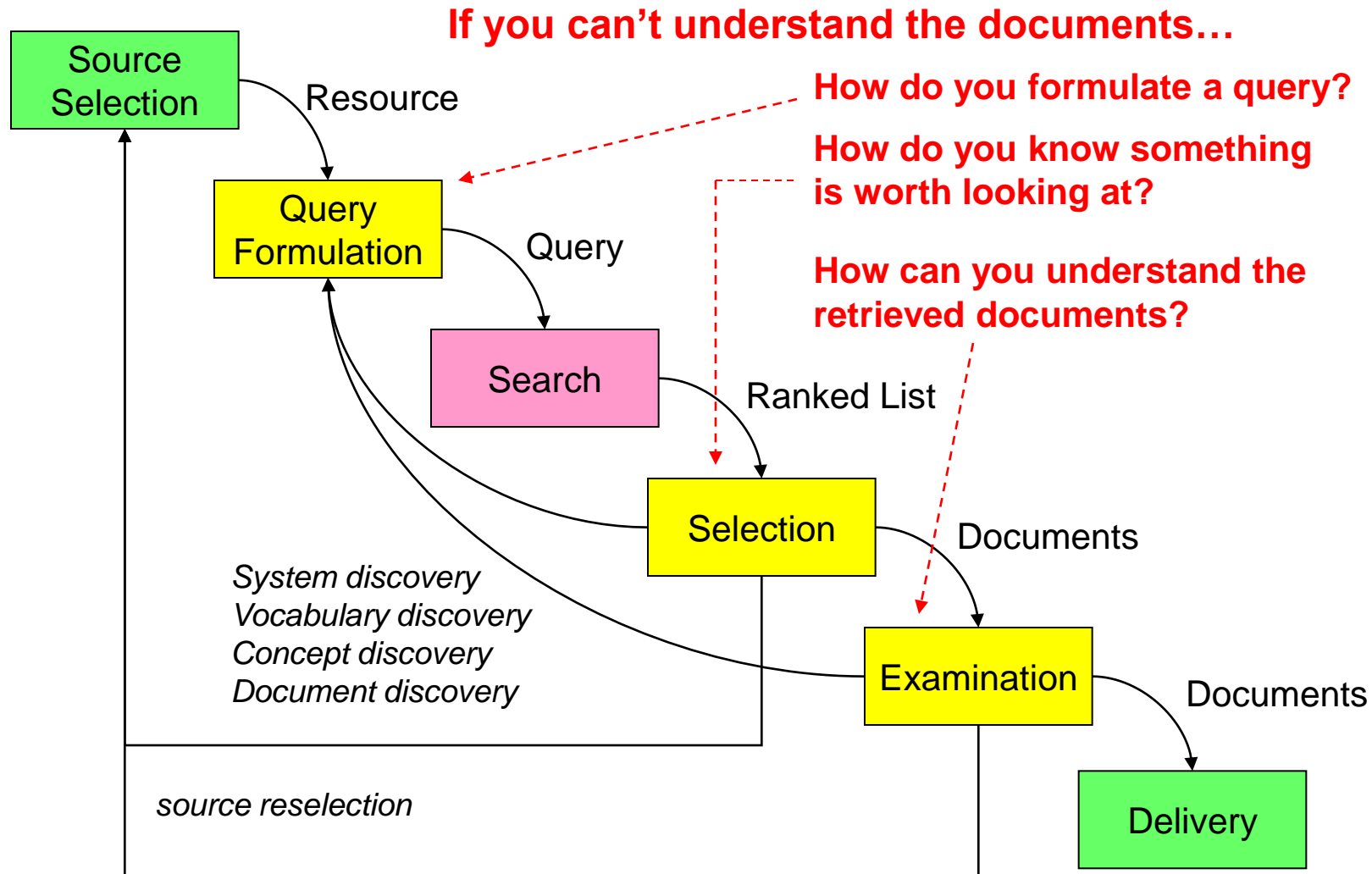
- For human understanding
- For automated processing



# A Little (Confusing) Vocabulary

- Multilingual document
  - Document containing more than one language
- Multilingual collection
  - Collection of documents in different languages
- Multilingual system
  - Can retrieve from a multilingual collection
- Cross-language system
  - Query in one language finds document in another
- Translingual system
  - Queries can find documents in any language

# The Information Retrieval Cycle



# Information Access

# Information Use

Translingual  
Search

Translingual  
Browsing

Translation

Select

Examine

Query

Document



# Early Work

- 1964 International Road Research
  - Multilingual thesauri
- 1970 SMART
  - Dictionary-based free-text cross-language retrieval
- 1978 ISO Standard 5964 (revised 1985)
  - Guidelines for developing multilingual thesauri
- 1990 Latent Semantic Indexing
  - Corpus-based free-text translingual retrieval

# Multilingual Thesauri

- Build a cross-cultural knowledge structure
  - Cultural differences influence indexing choices
- Use language-independent descriptors
  - Matched to language-specific lead-in vocabulary
- Three construction techniques
  - Build it from scratch
  - Translate an existing thesaurus
  - Merge monolingual thesauri

# Multilingual Information Access

## Information Science

### Information Retrieval

- Cross-Language Retrieval
- Indexing Languages
- Machine-Assisted Indexing

### Digital Libraries

- Multilingual Metadata

### Information Use

- International Information Flow
- Diffusion of Innovation

### Automatic Abstracting

## Artificial Intelligence

### Natural Language Processing

- Machine Translation
- Information Extraction
- Text Summarization

### Ontological Engineering

- Multilingual Ontologies

### Knowledge Discovery

- Textual Data Mining

### Machine Learning

## Other Fields

### Human-Computer Interaction

- Localization
- Information Visualization

### World-Wide Web

- Web Internationalization

### Speech Processing

- Topic Detection and Tracking

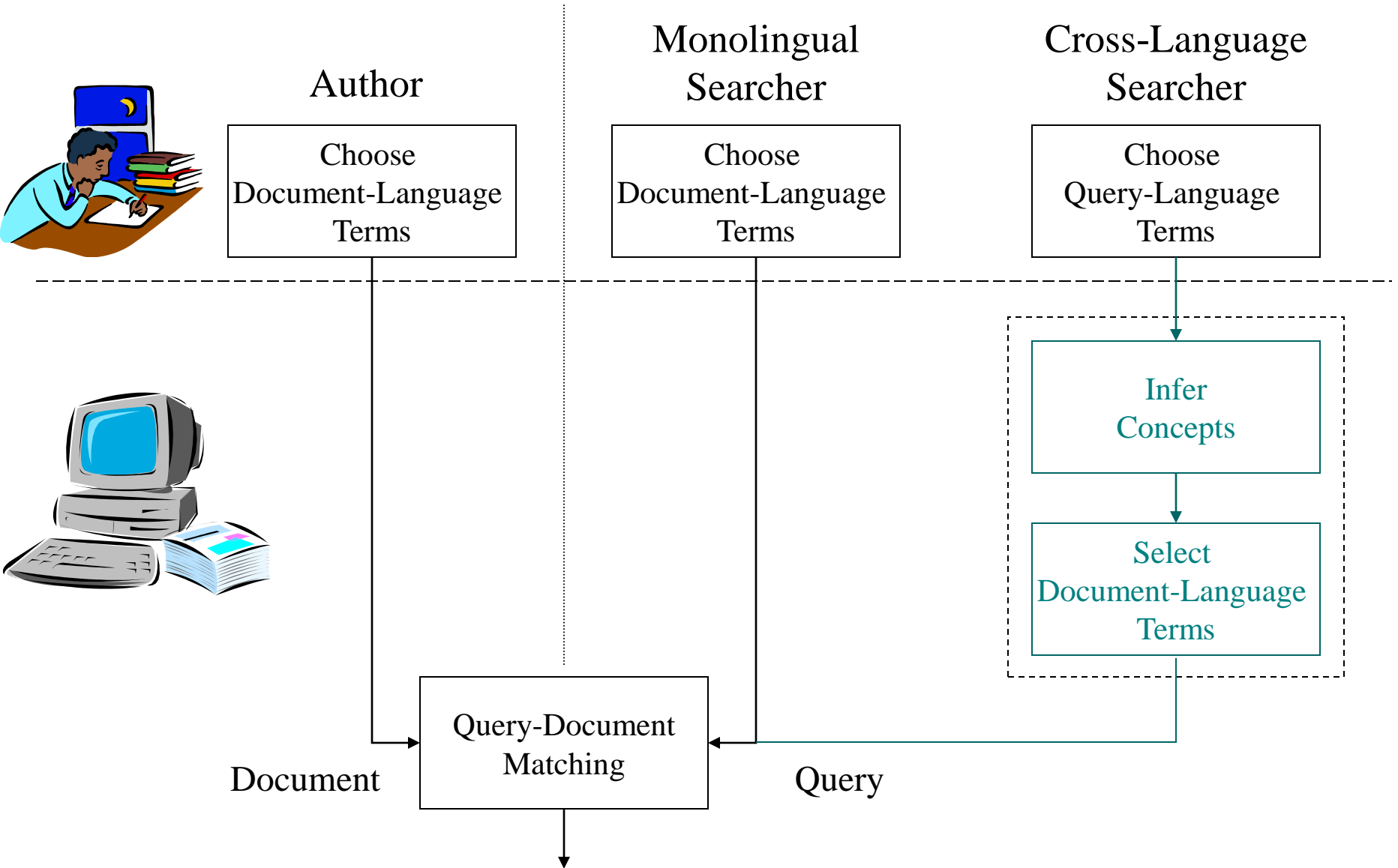
### Document Image Understanding

- Multilingual OCR

# Free Text CLIR

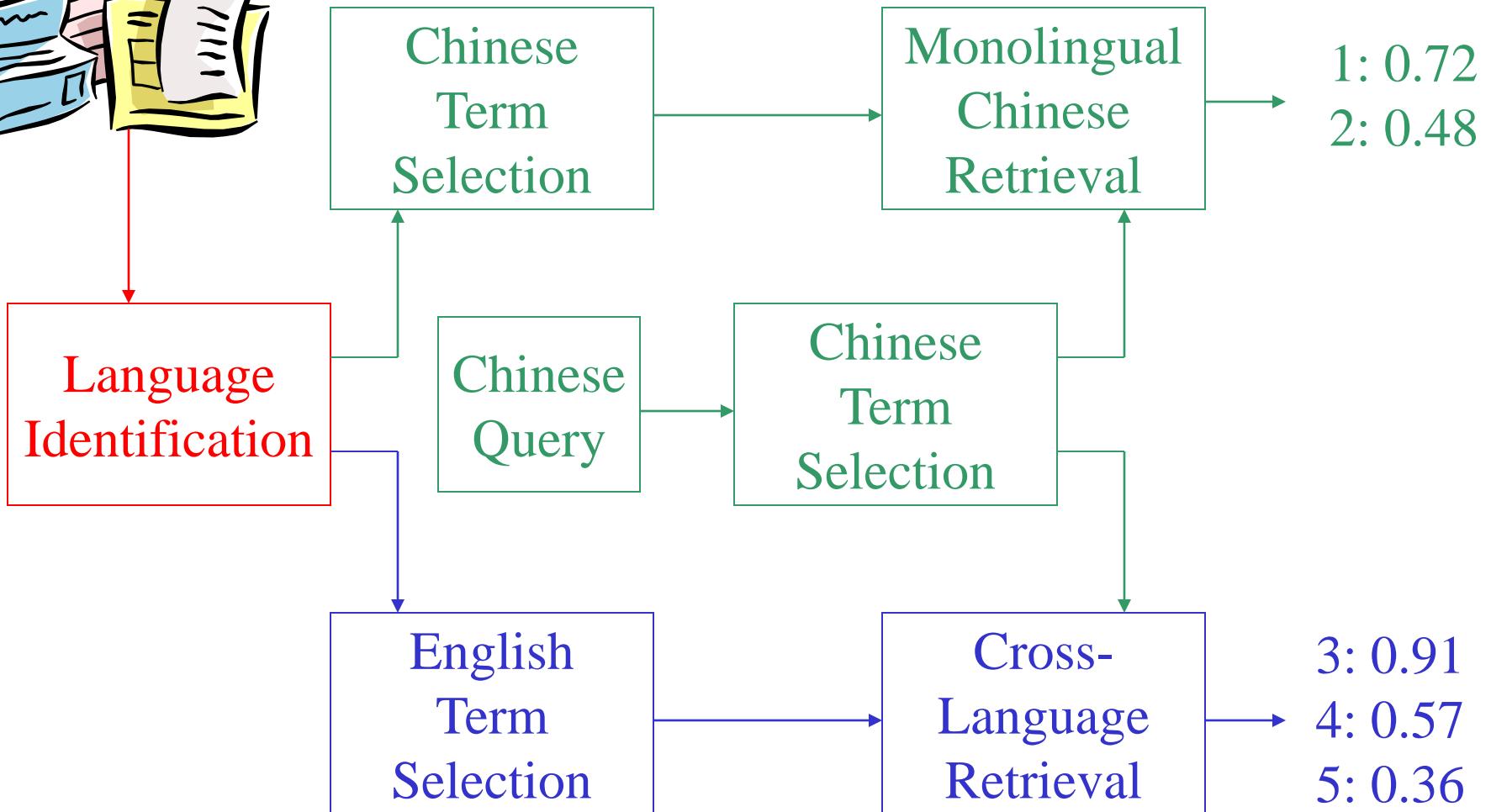
- What to translate?
  - Queries or documents
- Where to get translation knowledge?
  - Dictionary or corpus
- How to use it?

# The Search Process





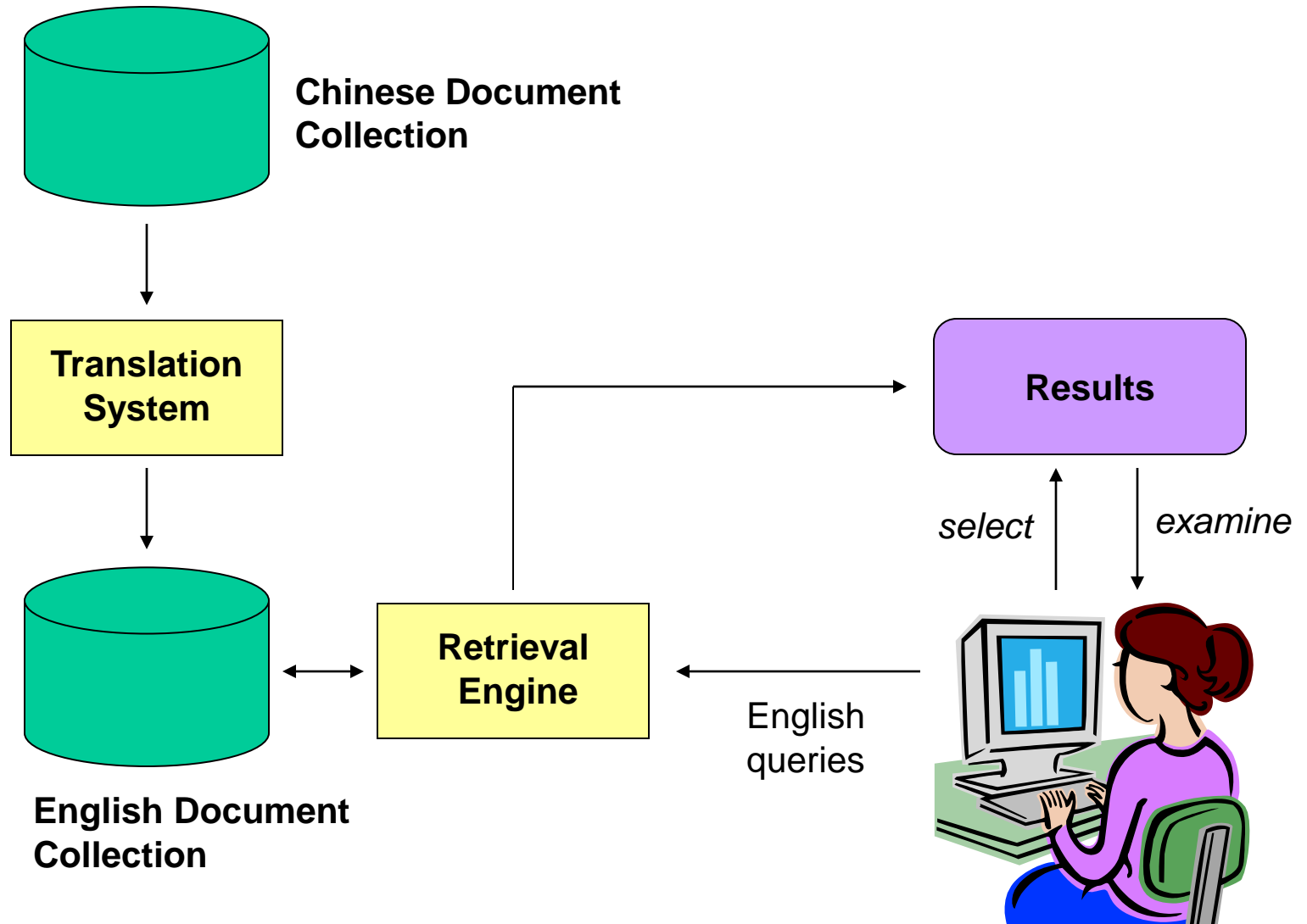
# Translingual Retrieval Architecture



# Evidence for Language Identification

- Metadata
  - Included in HTTP and HTML
- Word-scale features
  - Which dictionary gets the most hits?
- Subword features
  - Character n-gram statistics

# Query-Language IR

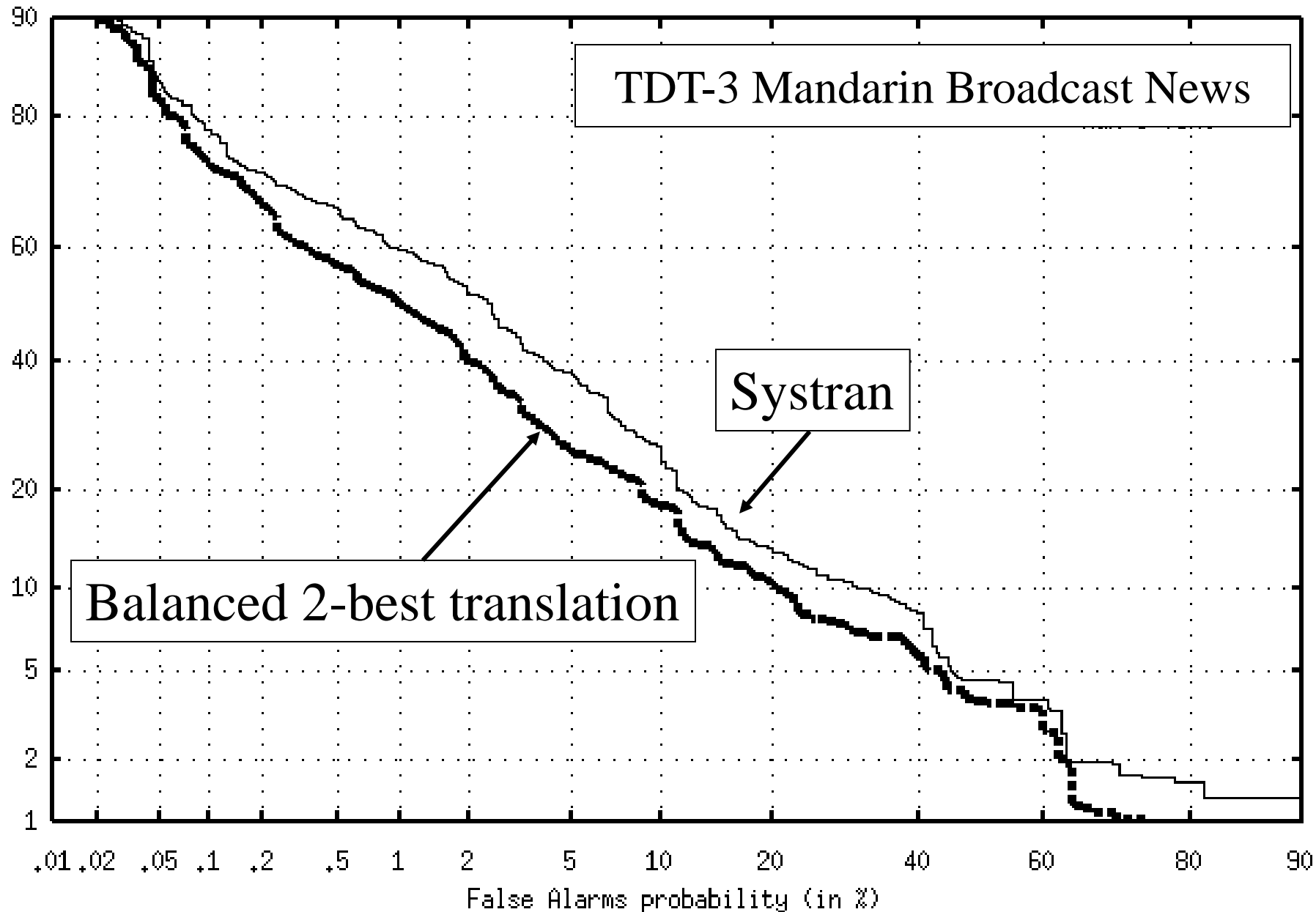


# Example: Modular use of MT

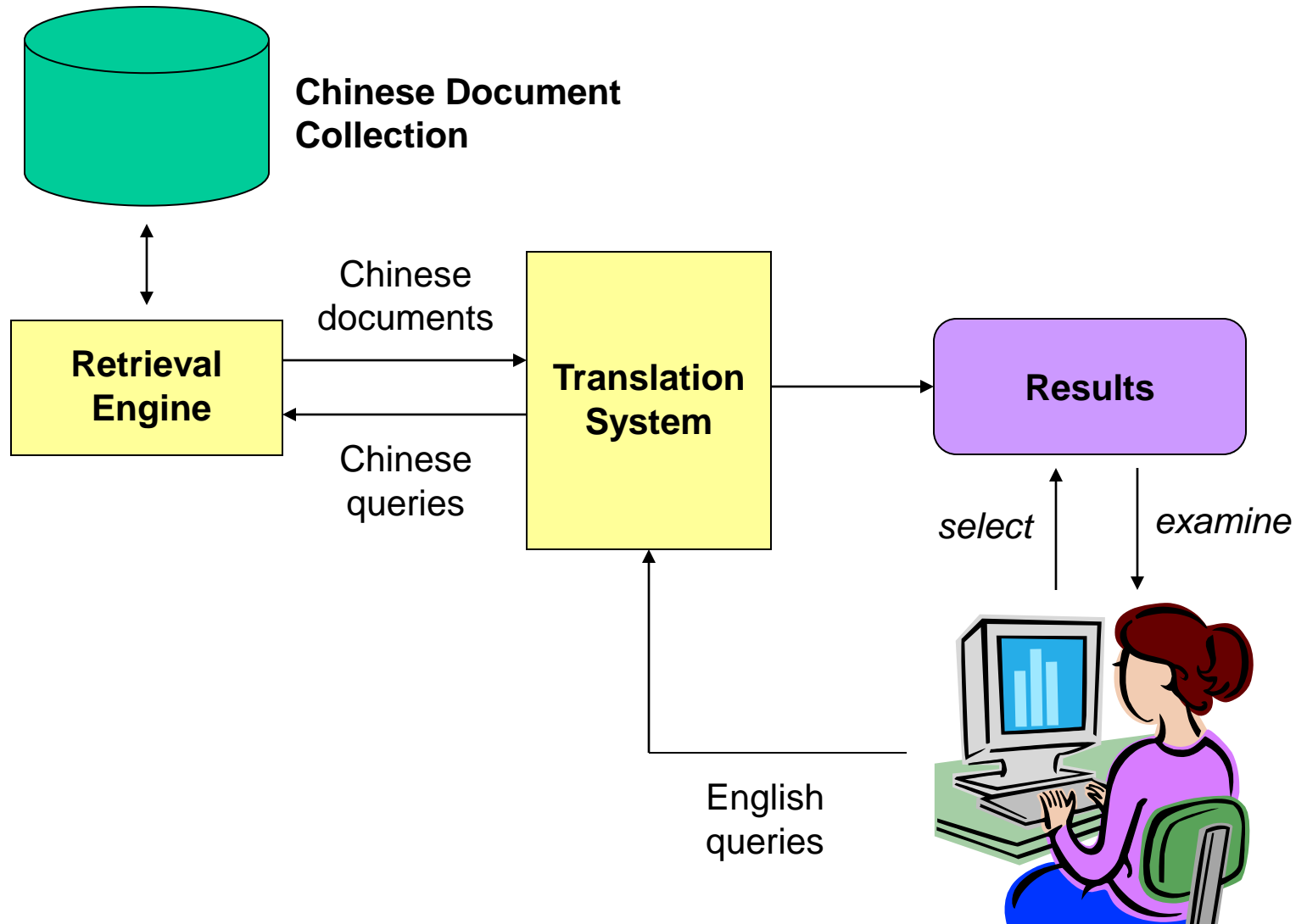
- Select a single query language
- Translate every document into that language
- Perform monolingual retrieval

# Is Machine Translation Enough?

Miss probability (in %)



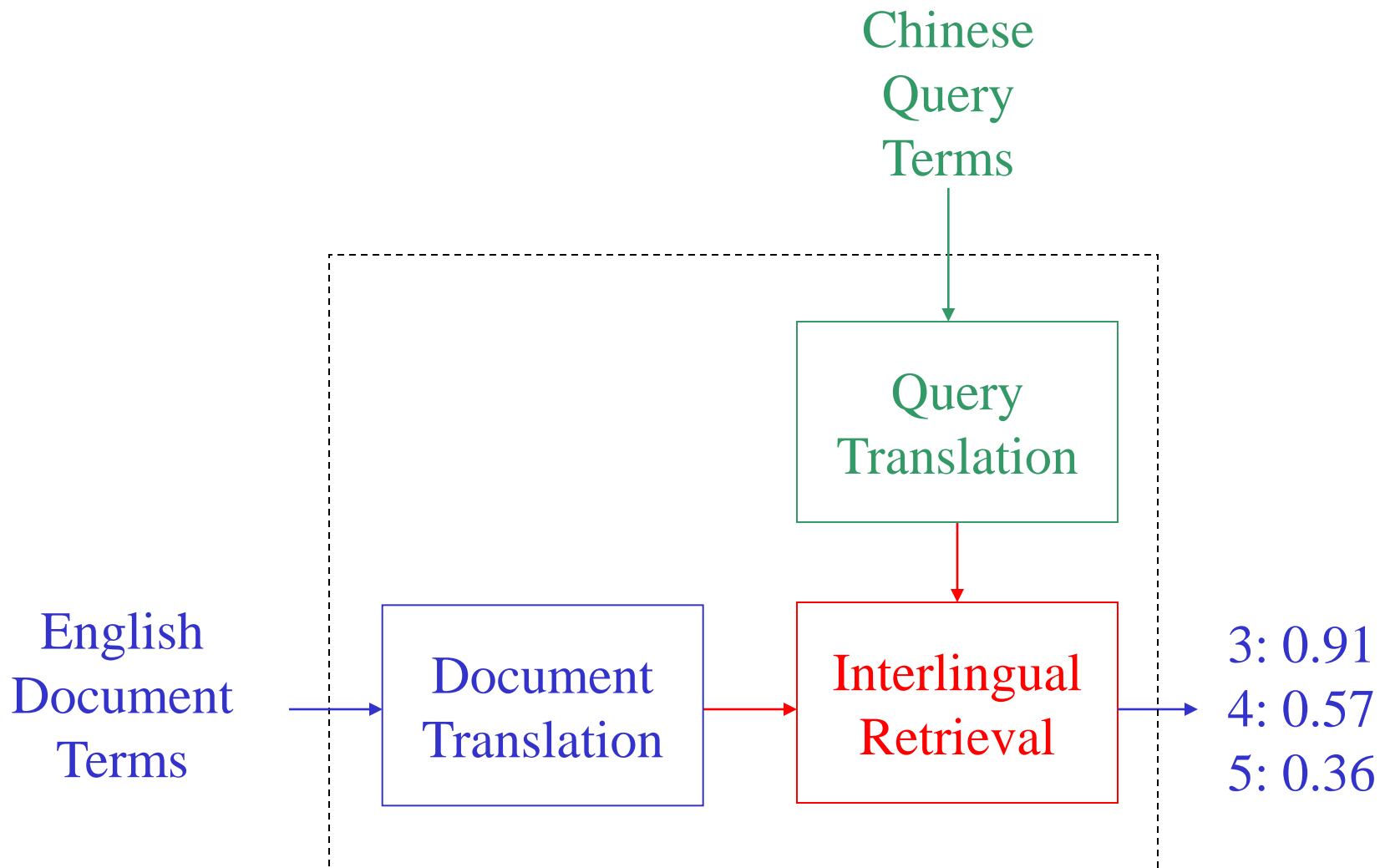
# Document-Language IR



# Query vs. Document Translation

- Query translation
  - Efficient for short queries (not relevance feedback)
  - Limited context for ambiguous query terms
- Document translation
  - Rapid support for interactive selection
  - Need only be done once (if query language is same)
- Merged query and document translation
  - Can produce better effectiveness than either alone

# Interlingual Retrieval





# Learning From Document Pairs

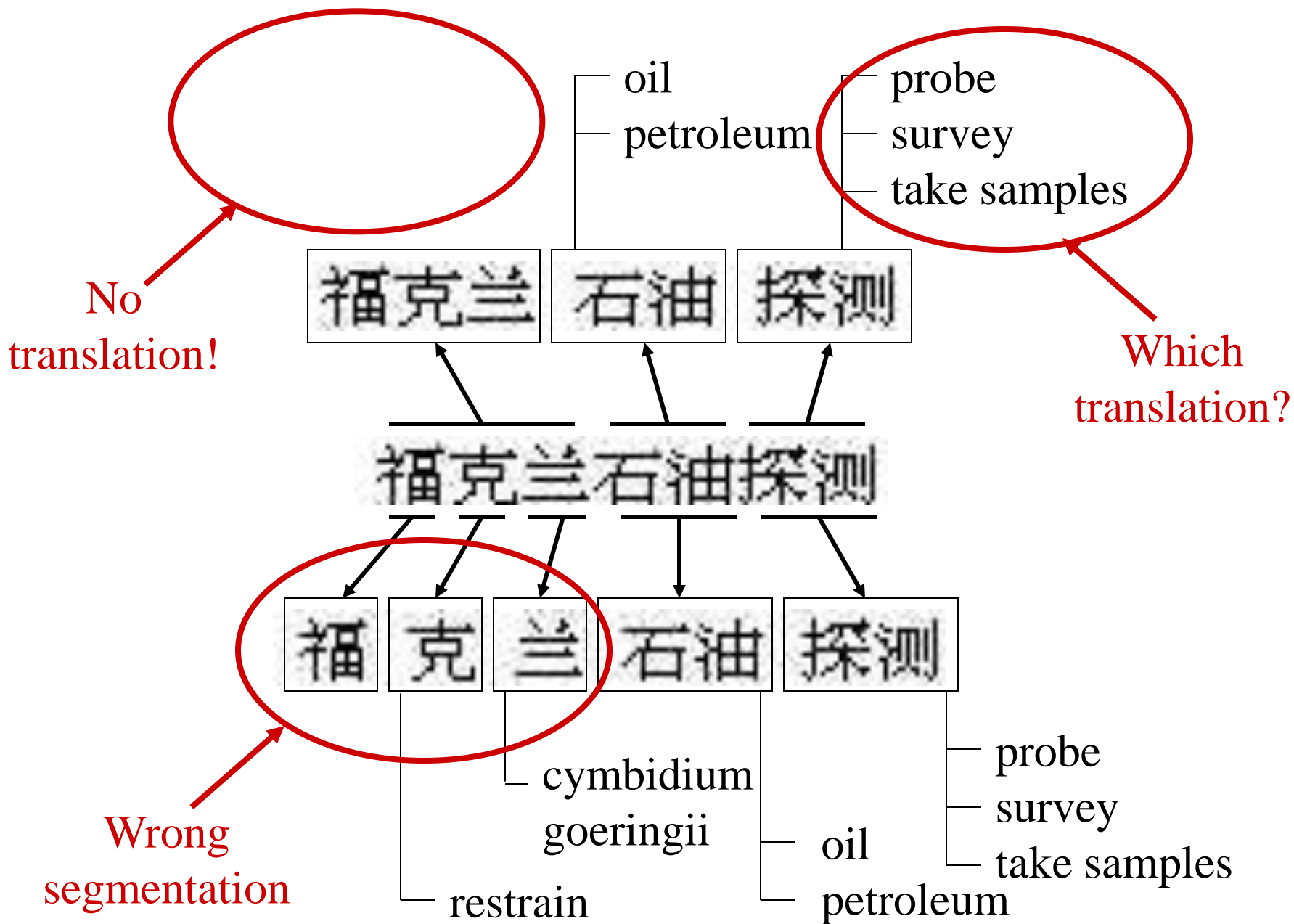
	English Terms					Spanish Terms			
	E1	E2	E3	E4	E5	S1	S2	S3	S4
Doc 1	4		2			2			1
Doc 2	8		4			4			2
Doc 3		2		2			2	1	
Doc 4		2	1				2		1
Doc 5	4				1	2		1	

# Generalized Vector Space Model

- “Term space” of each language is different
  - Document links define a common “document space”
- Describe documents based on the corpus
  - Vector of similarities to each corpus document
- Compute cosine similarity in document space
- Very effective in a within-domain evaluation

# Latent Semantic Indexing

- Cosine similarity captures noise with signal
  - Term choice variation and word sense ambiguity
- Signal-preserving dimensionality reduction
  - Conflates terms with similar usage patterns
    - Reduces term choice effect, even across languages
- Computationally expensive



# What's a “Term?”

- Granularity of a “term” depends on the task
  - Long for translation, more fine-grained for retrieval
- Phrases improve translation two ways
  - Less ambiguous than single words
  - Idiomatic expressions translate as a single concept
- Three ways to identify phrases
  - Semantic (e.g., appears in a dictionary)
  - Syntactic (e.g., parse as a noun phrase)
  - Co-occurrence (appear together unexpectedly often)

# Learning to Translate

- Lexicons
  - Phrase books, bilingual dictionaries, ...
- Large text collections
  - Translations (“parallel”)
  - Similar topics (“comparable”)
- Similarity
  - Similar pronunciation
- People

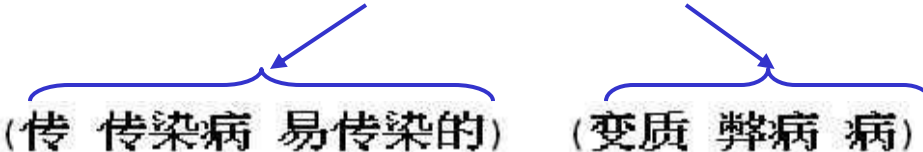
# Types of Lexical Resources

- Ontology
  - Organization of knowledge
- Thesaurus
  - Ontology specialized to support search
- Dictionary
  - Rich word list, designed for use by people
- Lexicon
  - Rich word list, designed for use by a machine
- Bilingual term list
  - Pairs of translation-equivalent terms

# Dictionary-Based Query Translation

Original query: El Nino and infectious diseases

Term selection: “El Nino” infectious diseases

Term translation:  (传 传染病 易传染的) (变质 弊病 病)

(Dictionary coverage: “El Nino” is not found)

Translation selection: (传染病 易传染的) 病

Query formulation:

Structure: OP1 (OP2 (传染病 易传染的) 病))



# Four-Stage Backoff

- Tralex might contain stems, surface forms, or some combination of the two.

## Document

mangez

surface form

mangez → mange

stem

mange

surface form

mangez → mange

stem

## Translation Lexicon

mangez - eat

surface form

mange - eats → eat

surface form

mangez → mange - eat

stem

mangent → mange - eat

stem

# Exploiting Part-of-Speech (POS)

- Constrain translations by part-of-speech
  - Requires POS tagger and POS-tagged lexicon
- Works well when queries are full sentences
  - Short queries provide little basis for tagging
- Constrained matching can hurt monolingual IR
  - Nouns in queries often match verbs in documents

# BM-25

$$\sum_{e \in Q} \left[ \log \frac{(N - df(e) + 0.5)}{(df(e) + 0.5)} \right] \left[ \frac{(2.2 * tf(e, d_k))}{(0.3 + 0.9 * \frac{dl(d_k)}{avdl} + tf(e, d_k))} \frac{8 * qtf(e)}{7 + qtf(e)} \right]$$

document frequency

term frequency

document length

# “Structured Queries”

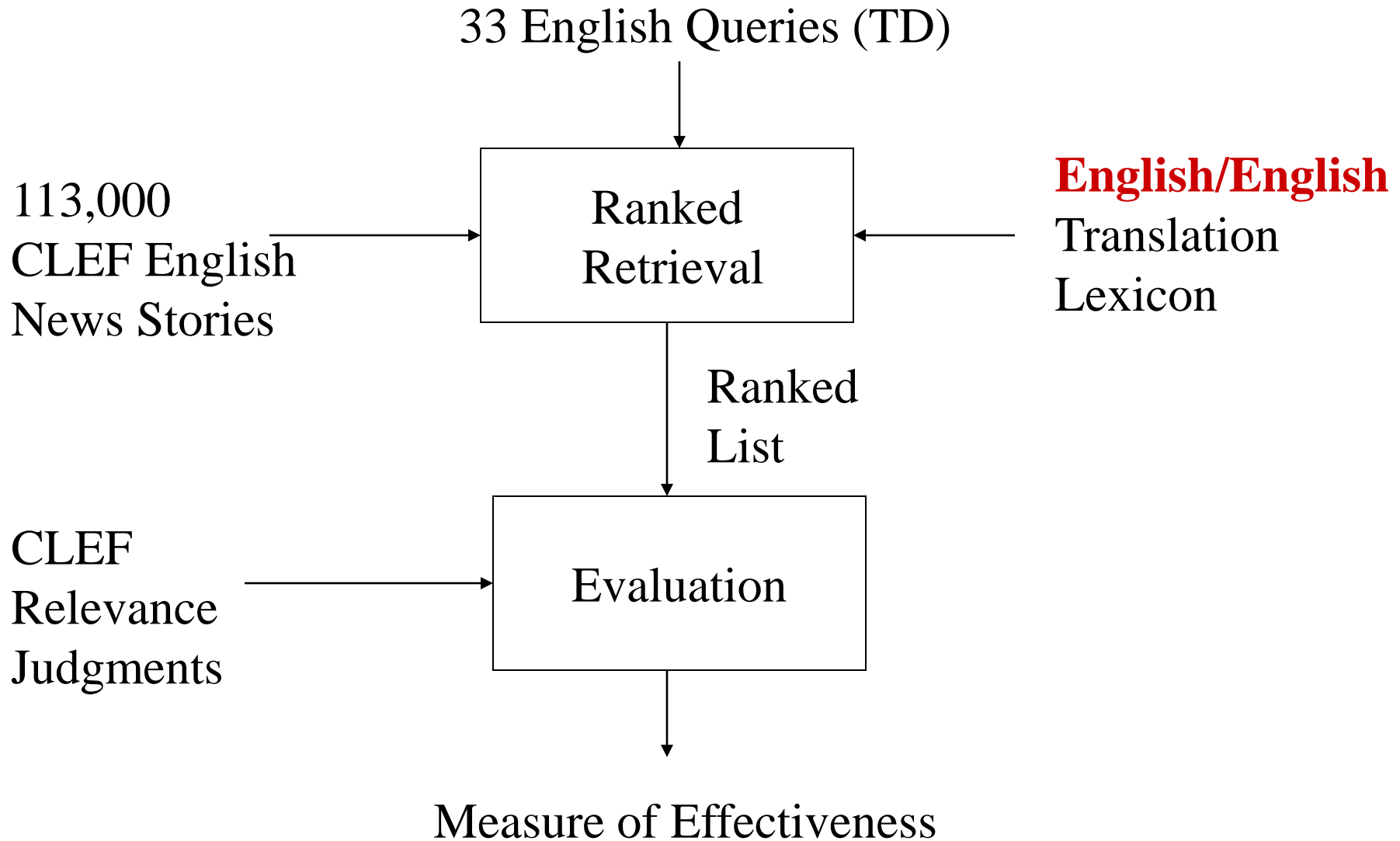
- Weight of term  $a$  in a document  $i$  depends on:
  - $TF(a,i)$ : Frequency of term  $a$  in document  $i$
  - $DF(a)$ : How many documents term  $a$  occurs in
- Build **pseudo-terms** from alternate translations
  - $TF(\mathbf{syn}(a,b),i) = TF(a,i) + TF(b,i)$
  - $DF(\mathbf{syn}(a,b)) = |\{\text{docs with } a\} \cup \{\text{docs with } b\}|$
- Downweight terms with any common translation
  - Particularly effective for long queries

# Computing Weights

- Unbalanced:  $\frac{1}{3} \left[ \frac{TF_1}{DF_1} + \frac{TF_2}{DF_2} + \frac{TF_3}{DF_3} \right]$ 
  - Overweights query terms that have many translations
- Balanced (#sum):  $\frac{1}{2} \left[ \frac{1}{2} \left( \frac{TF_1}{DF_1} + \frac{TF_2}{DF_2} \right) + \frac{TF_3}{DF_3} \right]$ 
  - Sensitive to rare translations
- Pirkola (#syn):  $\frac{1}{2} \left[ \frac{TF_1 + TF_2}{DF_1 \cup DF_2} + \frac{TF_3}{DF_3} \right]$ 
  - Deemphasizes query terms with any common translation

(Query Terms: 1: 传染病 2: 易传染的 3: 病 )

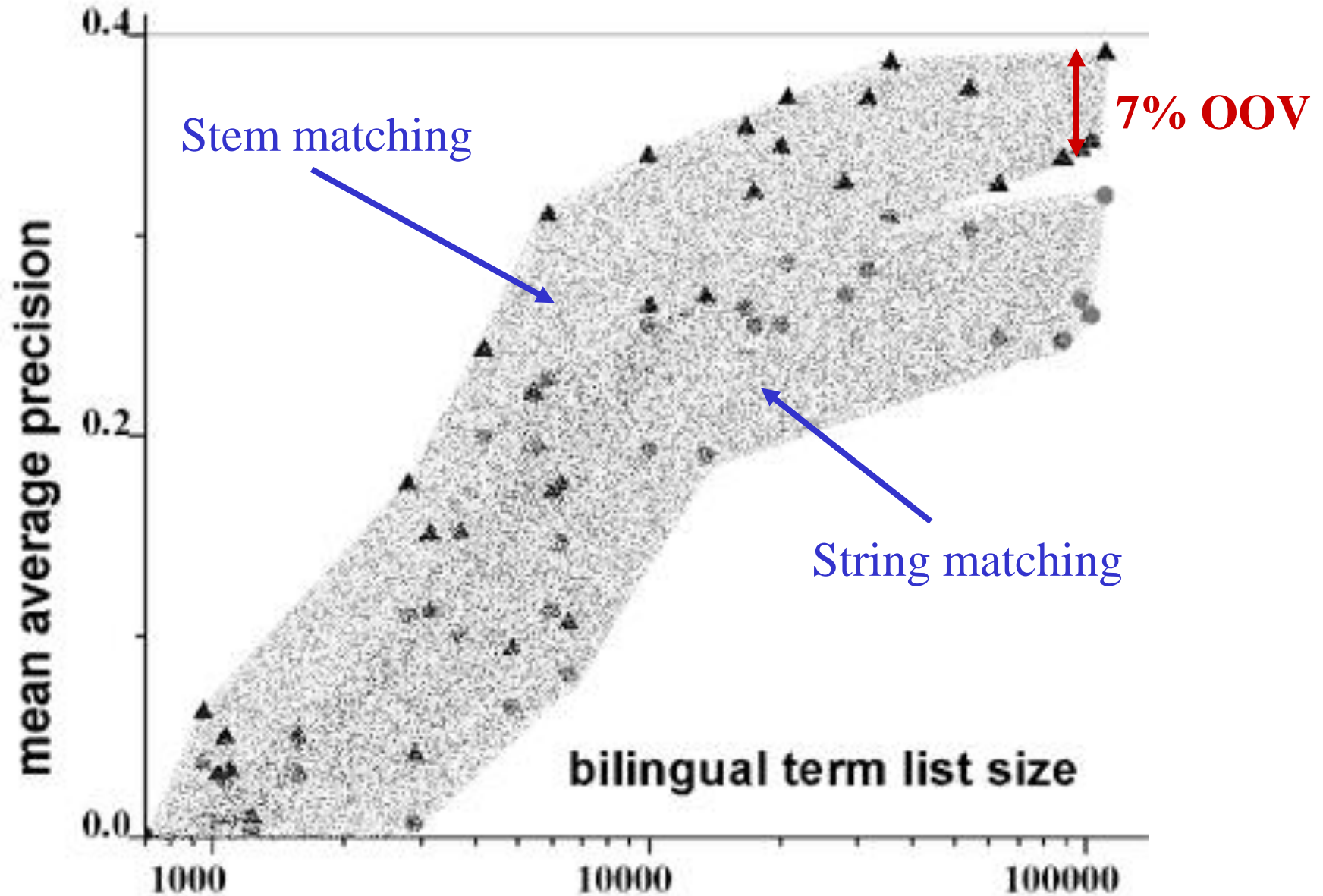
# Measuring Coverage Effects



# 35 Bilingual Term Lists

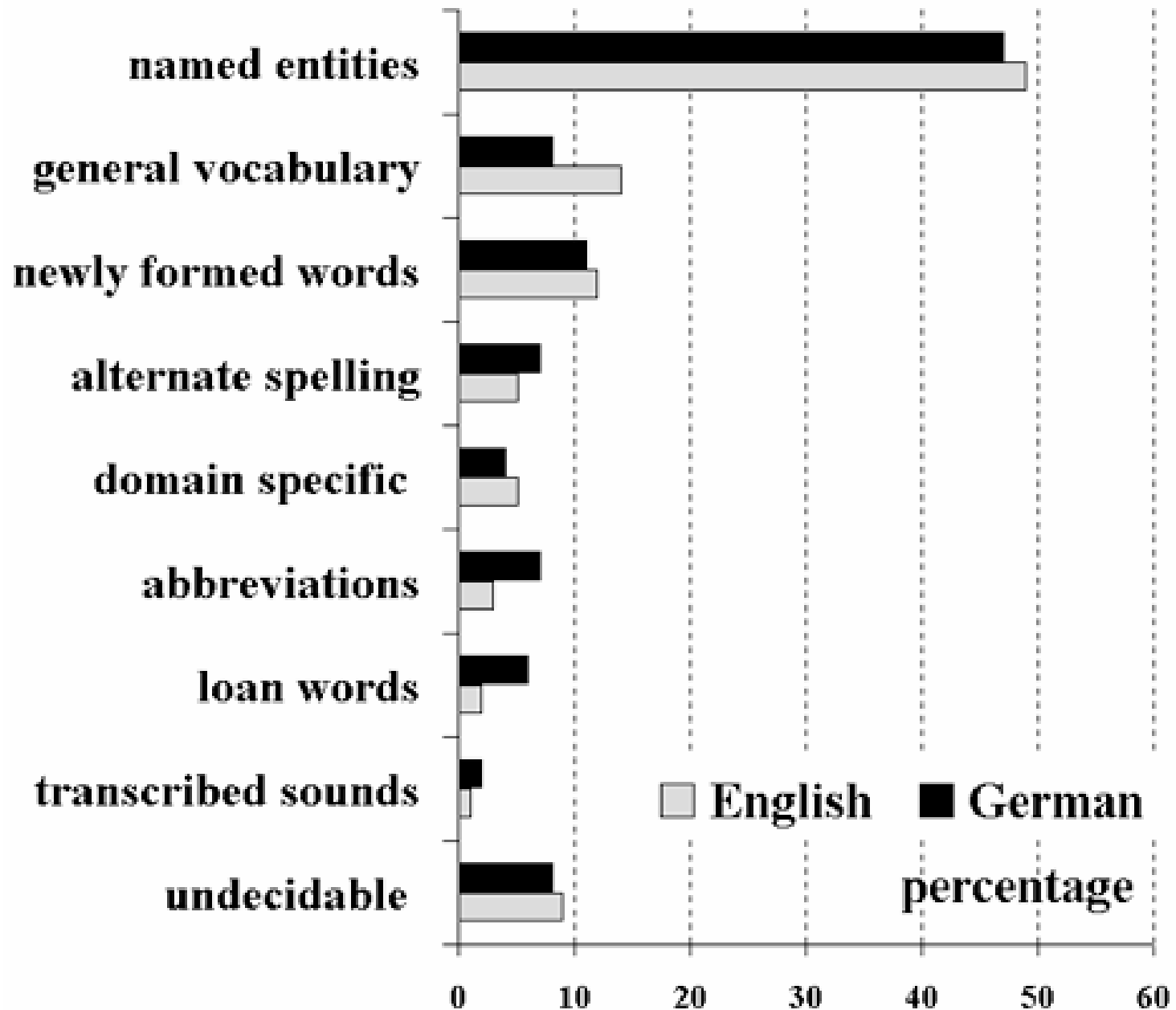
- Chinese (193, 111)
- German (103, 97, 89, 6)
- Hungarian (63)
- Japanese (54)
- Spanish (35, 21, 7)
- Russian (32)
- Italian (28, 13, 5)
- French (20, 17, 3)
- Esperanto (17)
- Swedish (10)
- Dutch (10)
- Norwegian (6)
- Portuguese (6)
- Greek (5)
- Afrikaans (4)
- Danish (4)
- Icelandic (3)
- Finnish (3)
- Latin (2)
- Welsh (1)
- Indonesian (1)
- Old English (1)
- Swahili (1)
- Eskimo (1)

# Size Effect

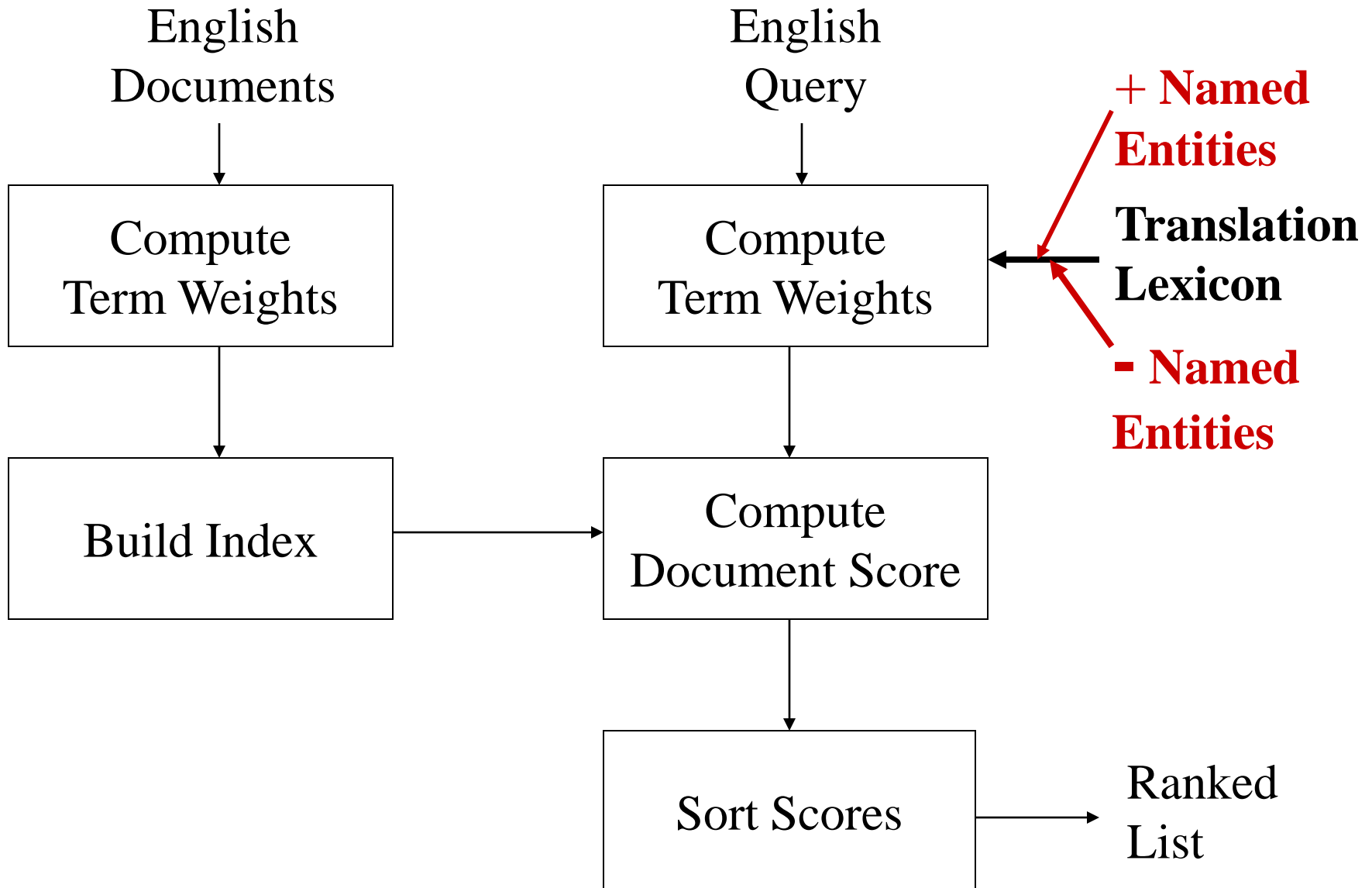


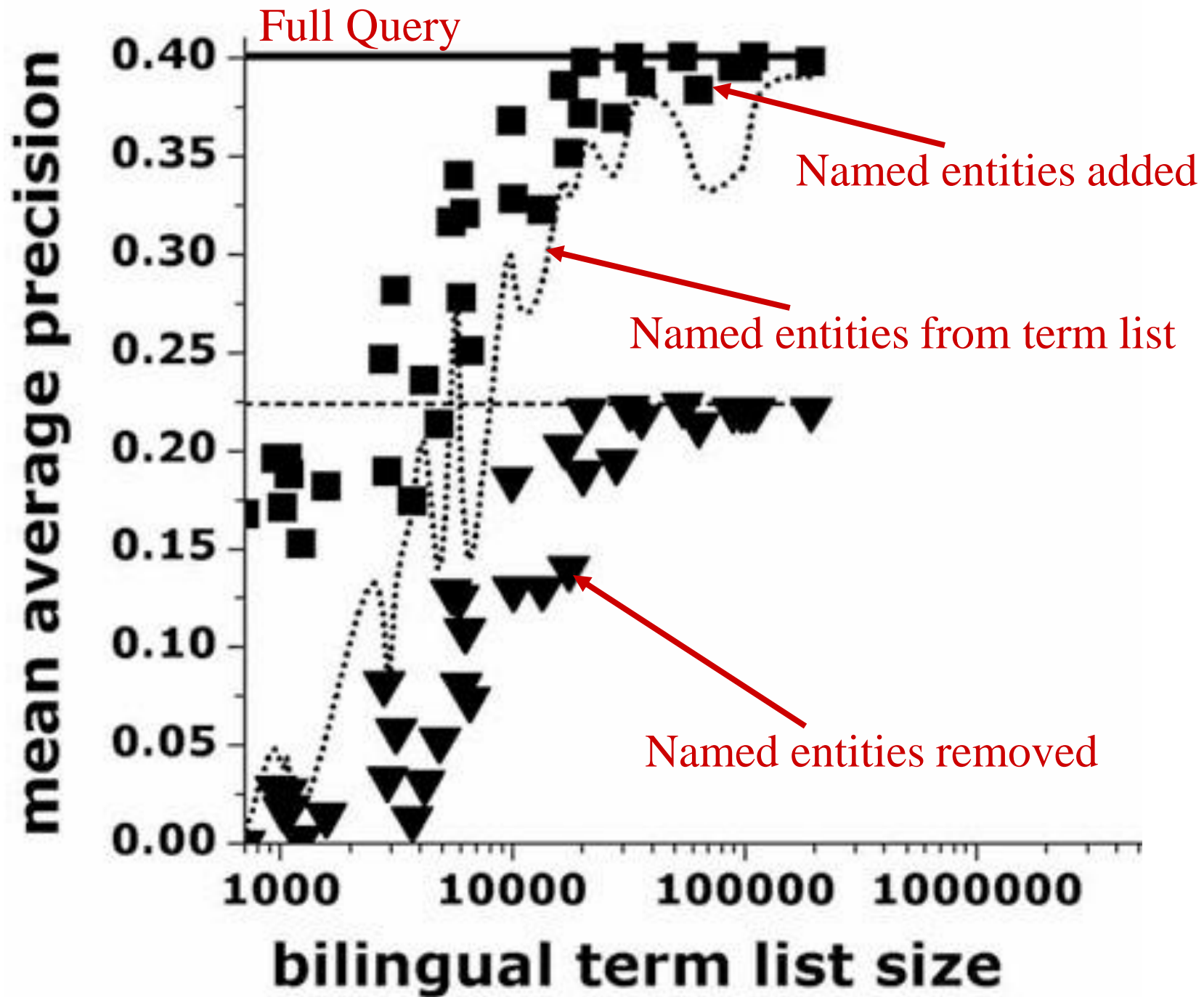


# Out-of-Vocabulary Distribution



# Measuring Named Entity Effect







Hieroglyphic

Egyptian Demotic

Greek

Fragment of the Rosetta Stone showing Hieroglyphic and Egyptian Demotic text. The Hieroglyphic text is at the top, and the Egyptian Demotic text is below it. The text is arranged in columns.

Fragment of the Rosetta Stone showing Greek text. The text is arranged in columns.

# Types of Bilingual Corpora

- Parallel corpora: translation-equivalent pairs
  - Document pairs
  - Sentence pairs
  - Term pairs
- Comparable corpora: topically related
  - Collection pairs
  - Document pairs

# Exploiting Parallel Corpora

- Automatic acquisition of translation lexicons
- Statistical machine translation
- Corpus-guided translation selection
- Document-linked techniques

# Some Modern Rosetta Stones

- News:
  - DE-News (German-English)
  - Hong-Kong News, Xinhua News (Chinese-English)
- Government:
  - Canadian Hansards (French-English)
  - Europarl (Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portugese, Spanish, Swedish)
  - UN Treaties (Russian, English, Arabic, ...)
- Religion
  - Bible, Koran, Book of Mormon

# Parallel Corpus

- Example from DE-News (8/1/1996)

**English:** Diverging opinions about planned tax reform

**German:** Unterschiedliche Meinungen zur geplanten Steuerreform

**English:** The discussion around the envisaged major tax reform continues .

**German:** Die Diskussion um die vorgesehene grosse Steuerreform dauert an .

**English:** The FDP economics expert , Graf Lambsdorff , today came out in favor of advancing the enactment of significant parts of the overhaul , currently planned for 1999 .

**German:** Der FDP - Wirtschaftsexperte Graf Lambsdorff sprach sich heute dafuer aus , wesentliche Teile der fuer 1999 geplanten Reform vorzuziehen .



# Word-Level Alignment

## English

Diverging opinions about planned tax reform

Unterschiedliche Meinungen zur geplanten Steuerreform

## German

## English

Madam President , I had asked the administration ...

Señora Presidenta, había pedido a la administración del Parlamento ...

## Spanish

# A Translation Model

- From word-aligned bilingual text, we induce a translation model

$$p(f_i | e) \quad \text{where,} \quad \sum_{f_i} p(f_i | e) = 1$$

- Example:

$$p(\text{探测} | \text{survey}) = 0.4$$

$$p(\text{试探} | \text{survey}) = 0.3$$

$$p(\text{测量} | \text{survey}) = 0.25$$

$$p(\text{样品} | \text{survey}) = 0.05$$

# Using Multiple Translations

- Weighted Structured Query Translation
  - Takes advantage of multiple translations and translation probabilities
- $TF$  and  $DF$  of query term  $e$  are computed using  $TF$  and  $DF$  of its translations:

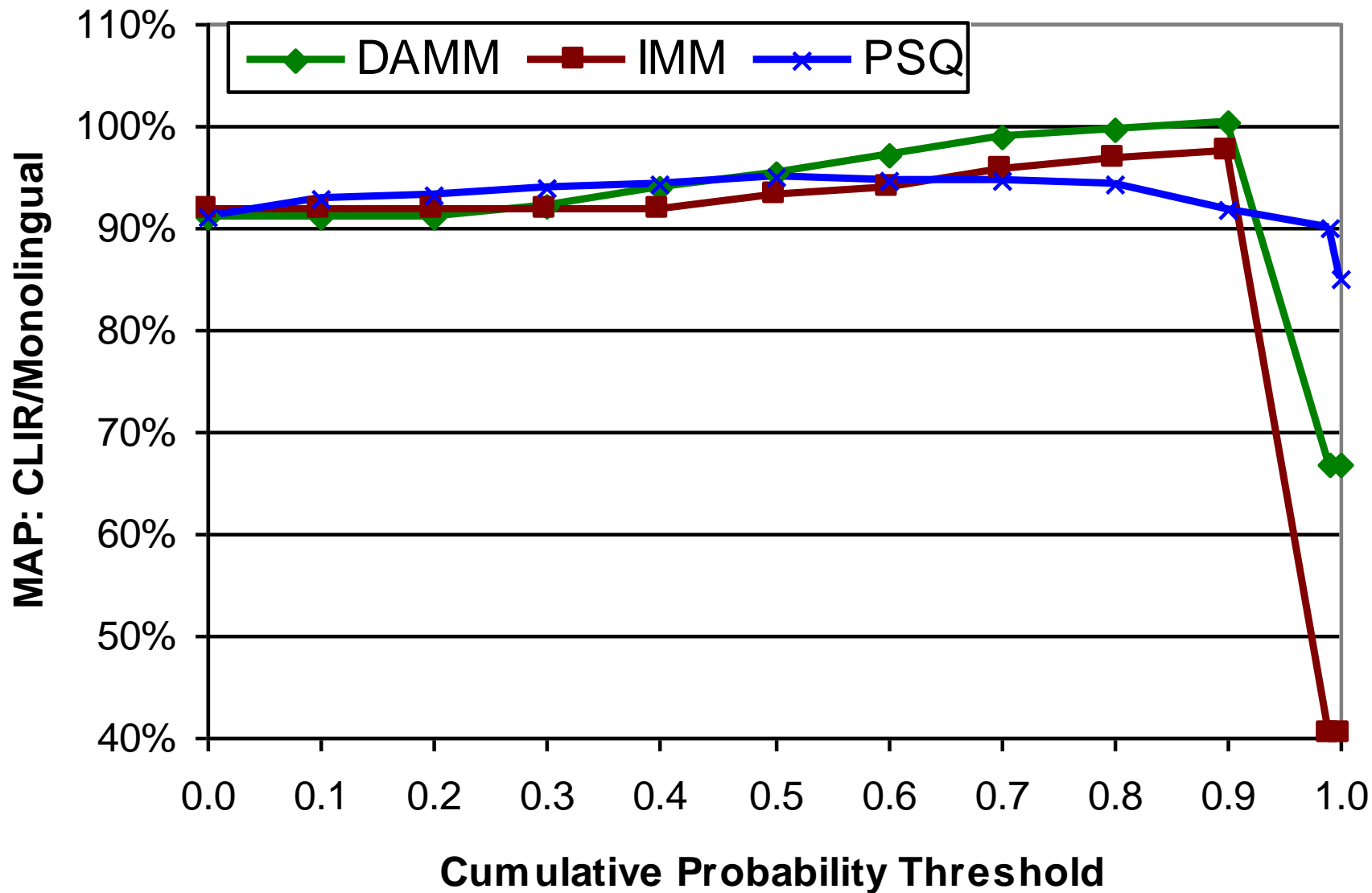
$$TF(e, D_k) = \sum_{f_i} p(f_i | e) \times TF(f_i, D_k)$$

$$DF(e) = \sum_{f_i} p(f_i | e) \times DF(f_i)$$

# Evaluating Corpus-Based Techniques

- Within-domain evaluation (upper bound)
  - Partition a bilingual corpus into training and test
  - Use the training part to tune the system
  - Generate relevance judgments for evaluation part
- Cross-domain evaluation (fair)
  - Use existing corpora and evaluation collections
  - No good metric for degree of domain shift

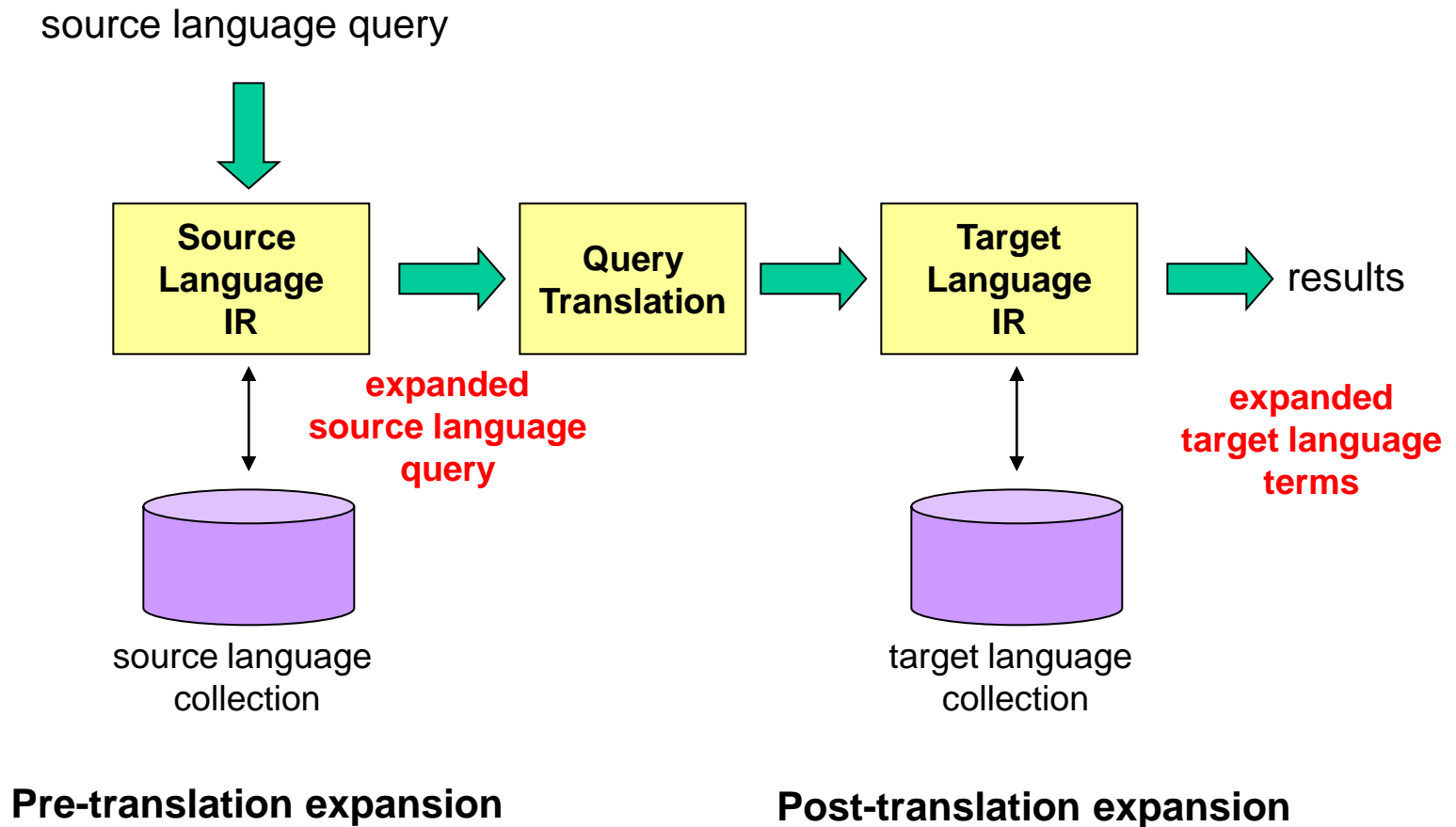
# Retrieval Effectiveness



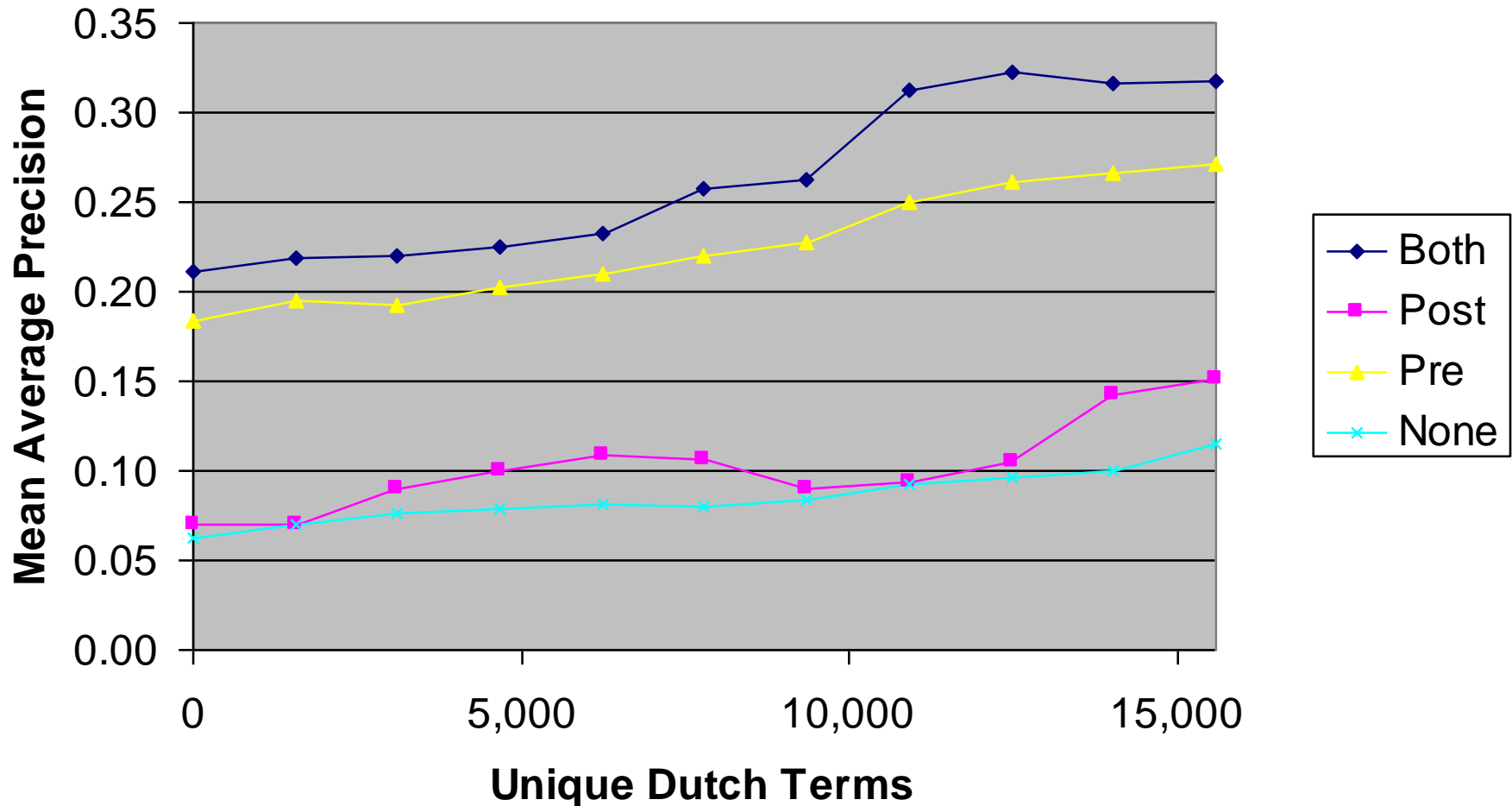
# Exploiting Comparable Corpora

- Blind relevance feedback
  - Existing CLIR technique + collection-linked corpus
- Lexicon enrichment
  - Existing lexicon + collection-linked corpus
- Dual-space techniques
  - Document-linked corpus

# Bilingual Query Expansion



# Query Expansion Effect

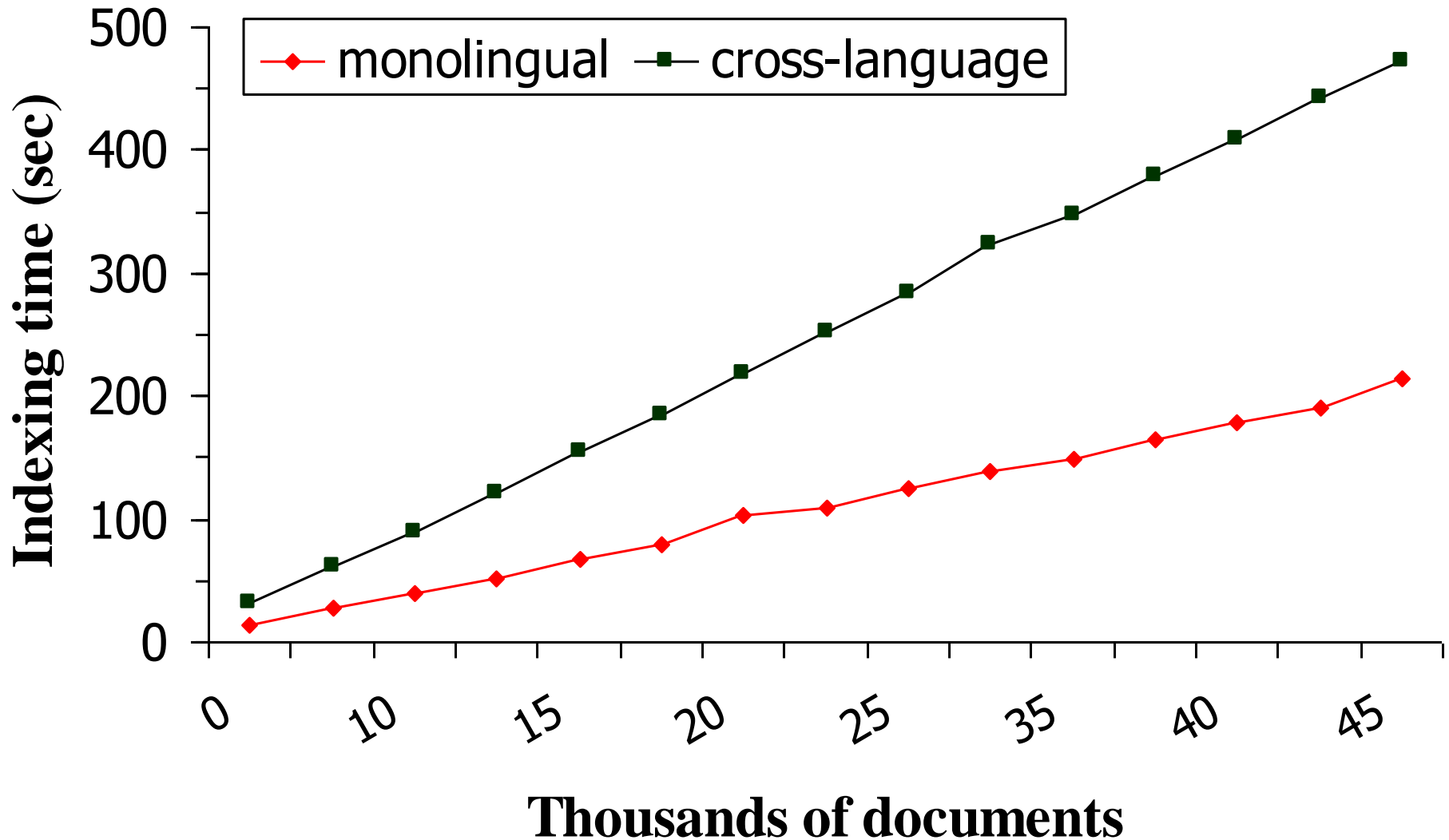




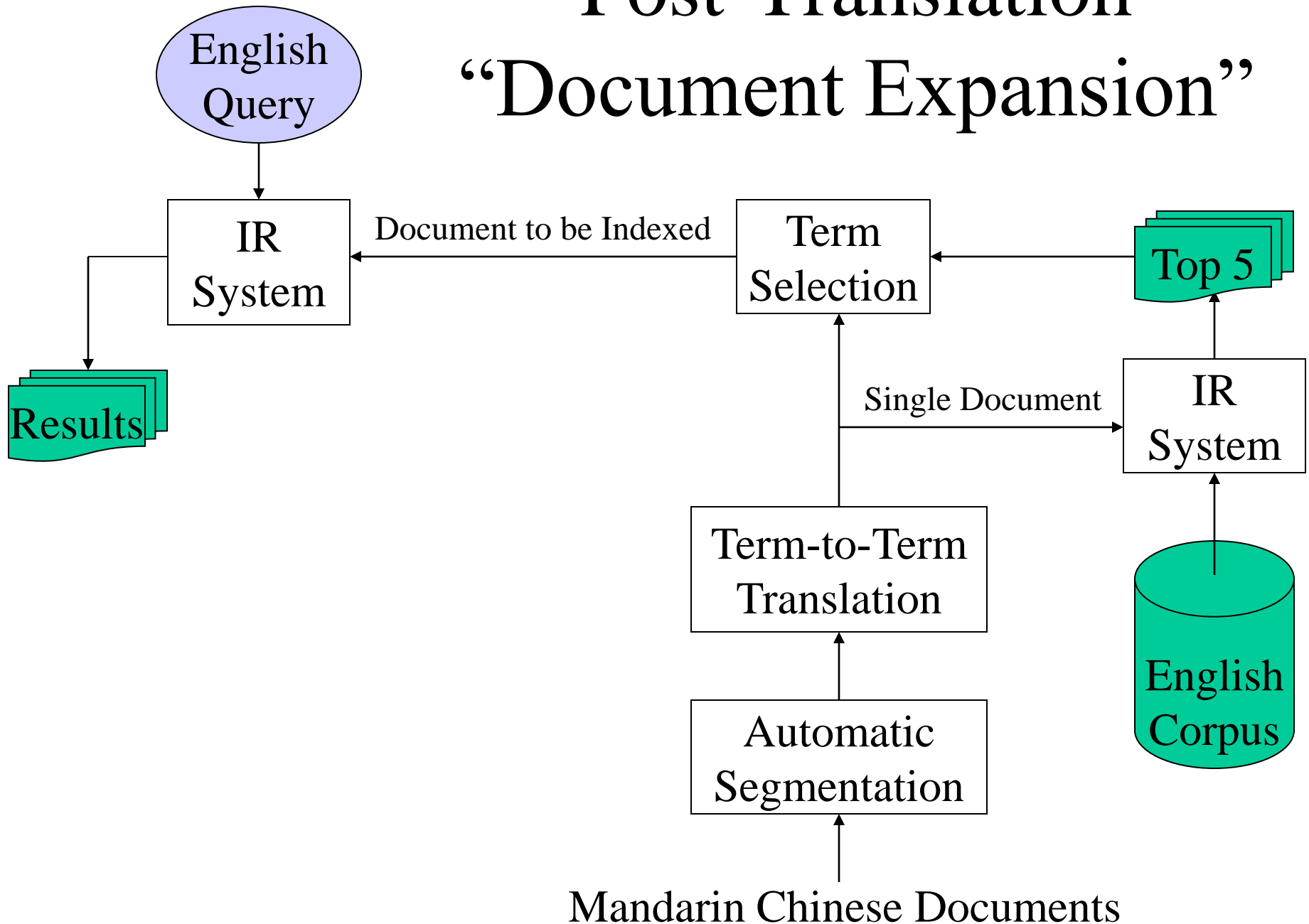
# Blind Relevance Feedback

- Augment a representation with related terms
  - Find related documents, extract distinguishing terms
- Multiple opportunities:
  - Before doc translation: Enrich the vocabulary
  - After doc translation: Mitigate translation errors
  - Before query translation: Improve the query
  - After query translation: Mitigate translation errors
- Short queries get the most dramatic improvement

# Indexing Time: Doc Translation



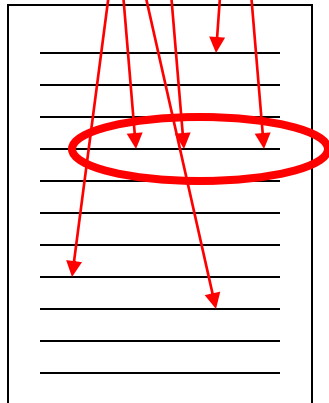
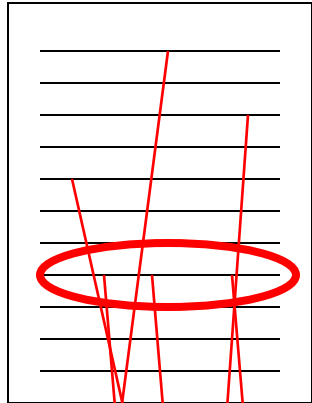
# Post-Translation “Document Expansion”



# Why Document Expansion Works

- Story-length objects provide useful context
- Ranked retrieval finds signal amid the noise
- Selective terms discriminate among documents
  - Enrich index with low DF terms from top documents
- Similar strategies work well in other applications
  - CLIR query translation
  - Monolingual spoken document retrieval

# Lexicon Enrichment



... Cross-Language Evaluation Forum ...

... Solto Extunifoc Tanixul Knadu ...

# Lexicon Enrichment

- Use a bilingual lexicon to align “context regions”
  - Regions with high coincidence of known translations
- Pair unknown terms with unmatched terms
  - Unknown: language A, not in the lexicon
  - Unmatched: language B, not covered by translation
- Treat the most surprising pairs as new translations

# Cognate Matching

- Dictionary coverage is inherently limited
  - Translation of proper names
  - Translation of newly coined terms
  - Translation of unfamiliar technical terms
- Strategy: model derivational translation
  - Orthography-based
  - Pronunciation-based

# Matching Orthographic Cognates

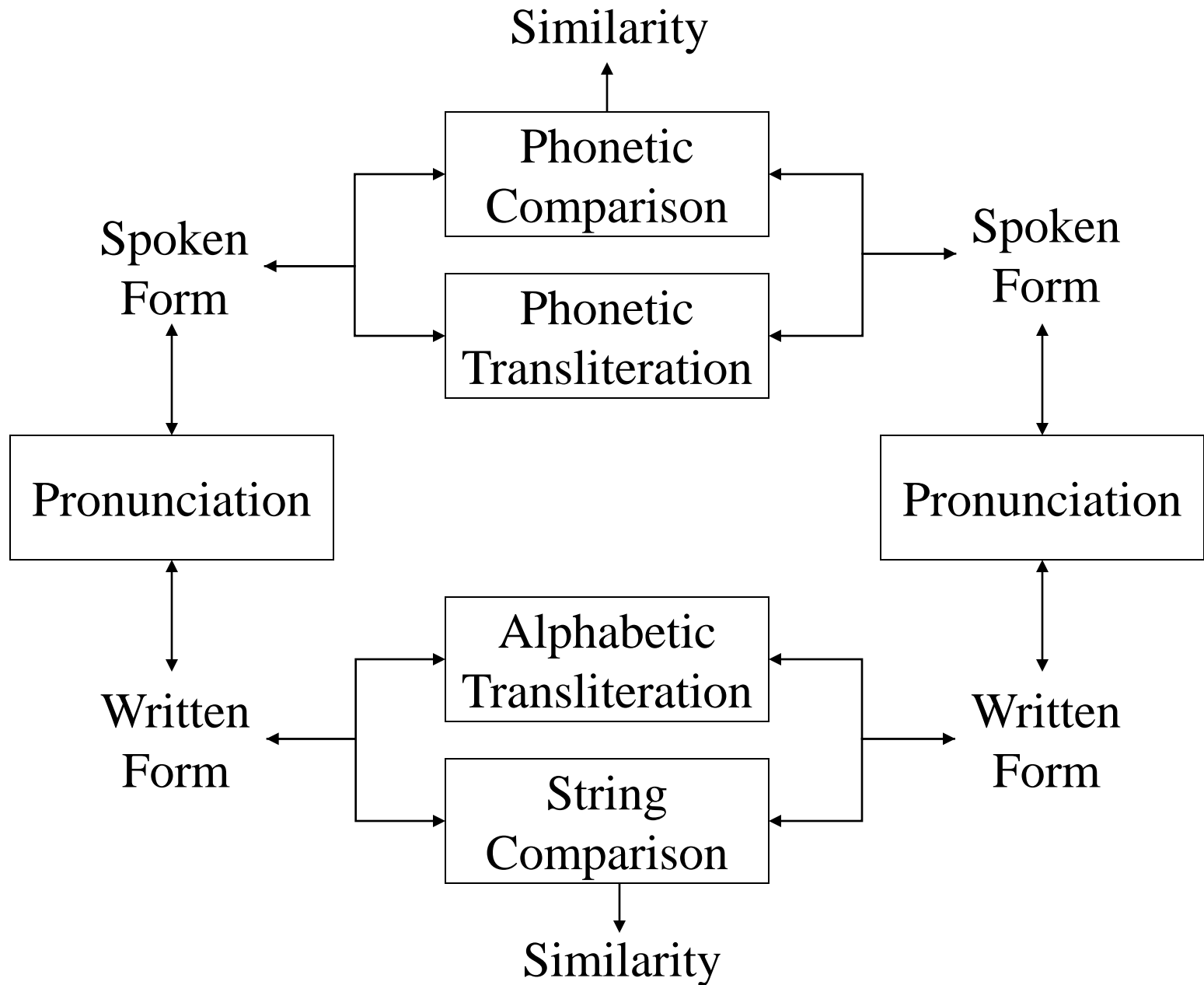
- Retain untranslatable words unchanged
  - Often works well between European languages
- Rule-based systems
  - Even off-the-shelf spelling correction can help!
- Character-level statistical MT
  - Trained using a set of representative cognates



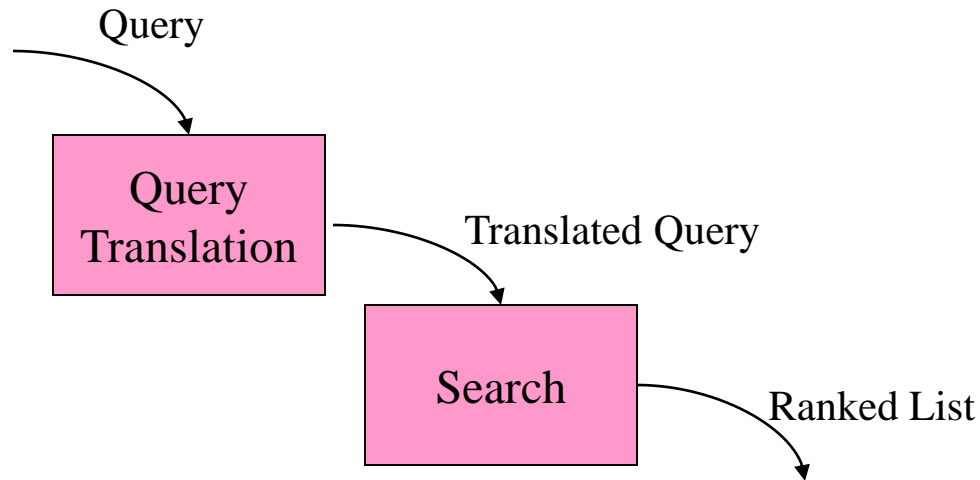
# Matching Phonetic Cognates

- Forward transliteration
  - Generate all potential transliterations
- Reverse transliteration
  - Guess source string(s) that produced a transliteration
- Match in phonetic space

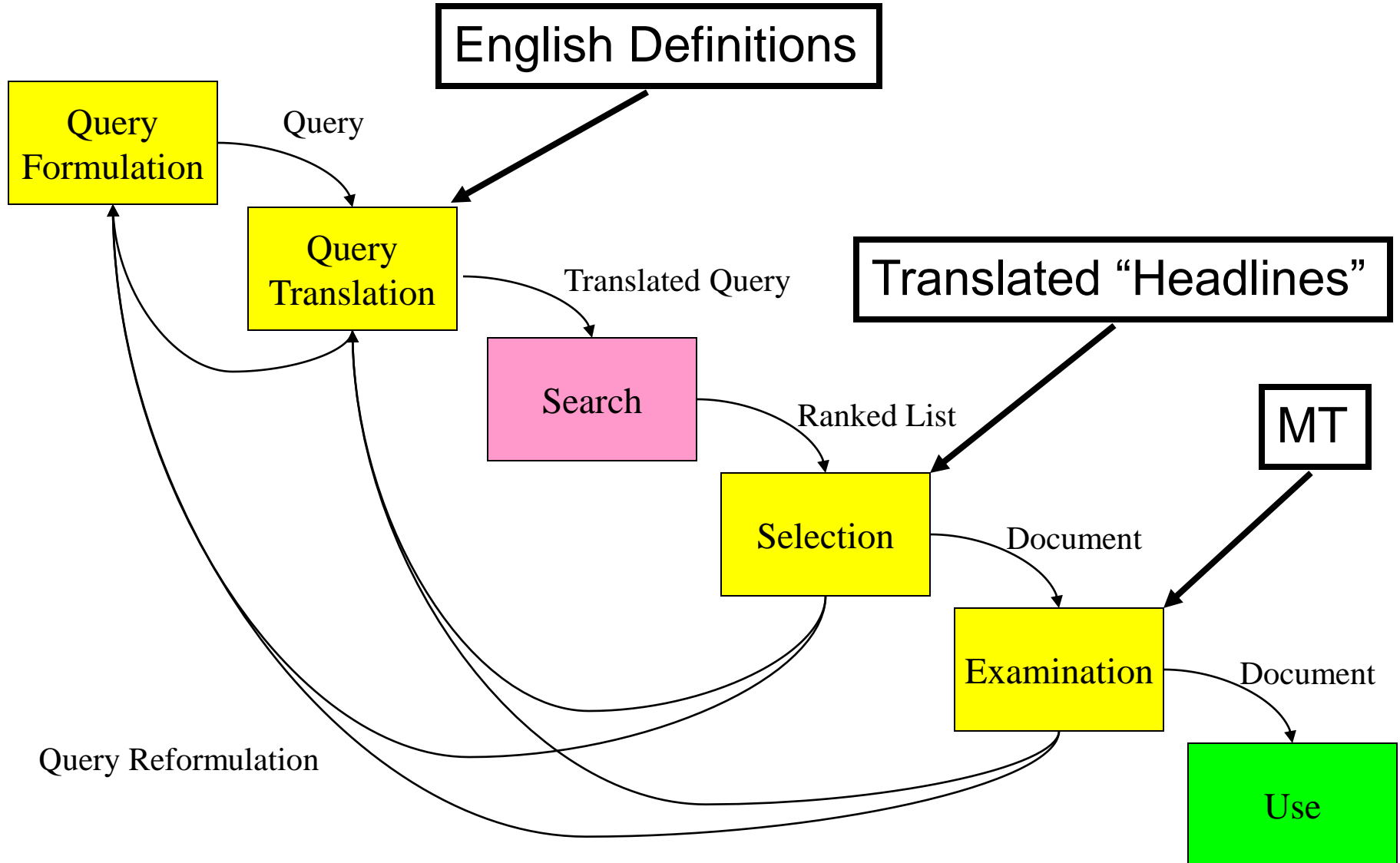
# Leveraging Cognates



# Cross-Language “Retrieval”



# Interactive Translingual Search



# Selection

- Goal: Provide information to support decisions
- May not require very good translations
  - e.g., Term-by-term title translation
- People can “read past” some ambiguity
  - May help to display a few alternative translations

**MIRACLE: Maryland Interactive Retrieval Advanced Cross-Language Engine**

**Collections   Configure   Display   Dictionaries   Help**

Look for: indian film and social and cultural impact

Search
Reset

**PREVIOUS QUERIES**  
**CURRENT QUERY**  
indian  
**film**  
bak.ckaahraaiyaaa  
chaikata  
failaahmaon  
jhailaahlaii  
kaaimarae kail raila  
sainaemaaa  
**social**

**FILM**

Select All
Deselect All

	Hindi	Probability	Synonym List	Sample Usage 1
<input checked="" type="checkbox"/>	bak.ckaahraaiy...		film	
<input checked="" type="checkbox"/>	chaikata		bacterial, sticky, of, film	
<input checked="" type="checkbox"/>	failaahmaon		designs, cartoon, film	
<input checked="" type="checkbox"/>	jhailaahlaii		peritonitis, lining, membrane, film	There is a #film# of ...
<input checked="" type="checkbox"/>	kaaimarae kail r...		film	
<input checked="" type="checkbox"/>	sainaemaaa		trip, matinee, cinema, be, film	The #film# now sho...

Search Again

1
2
3

.. of the organisation hand should not be but **cultural** , **social** and economic change of the car x ; - often violence and aggression of such an atmosphere where the violence ..... to people were killed . as far as **indian** society is concerned over the past few years in the violence to protest the non-violence to the **social** life of the largest ..... of the decline was . in fact , **social** violence of the traditional x ( ways violence in which a new look to have come to his imagination perhaps a was ...  
/data/mt/hindi/HTTP/www.bhaskar.com/050999/form.htm

.. e try e photo gallery e literature and **culture** e religion e e future / calendar the main page ' devdas ' oscar that atal most hindi **films** , the weary ..... prime minister atal bihari vajpayee , wiezacker hindi **film** industry news taken . he said that most **films** boring are " devdas ' them great like i . vajpayee expressed the ..... successful . vajpayee on tuesday , the telugu **film** of the giant bowel rao on the life of based monographs blackwemm dr. ' to issue on the spot programme to address ...  
14HHa\_id=838900\_pda=6/15/2002\_

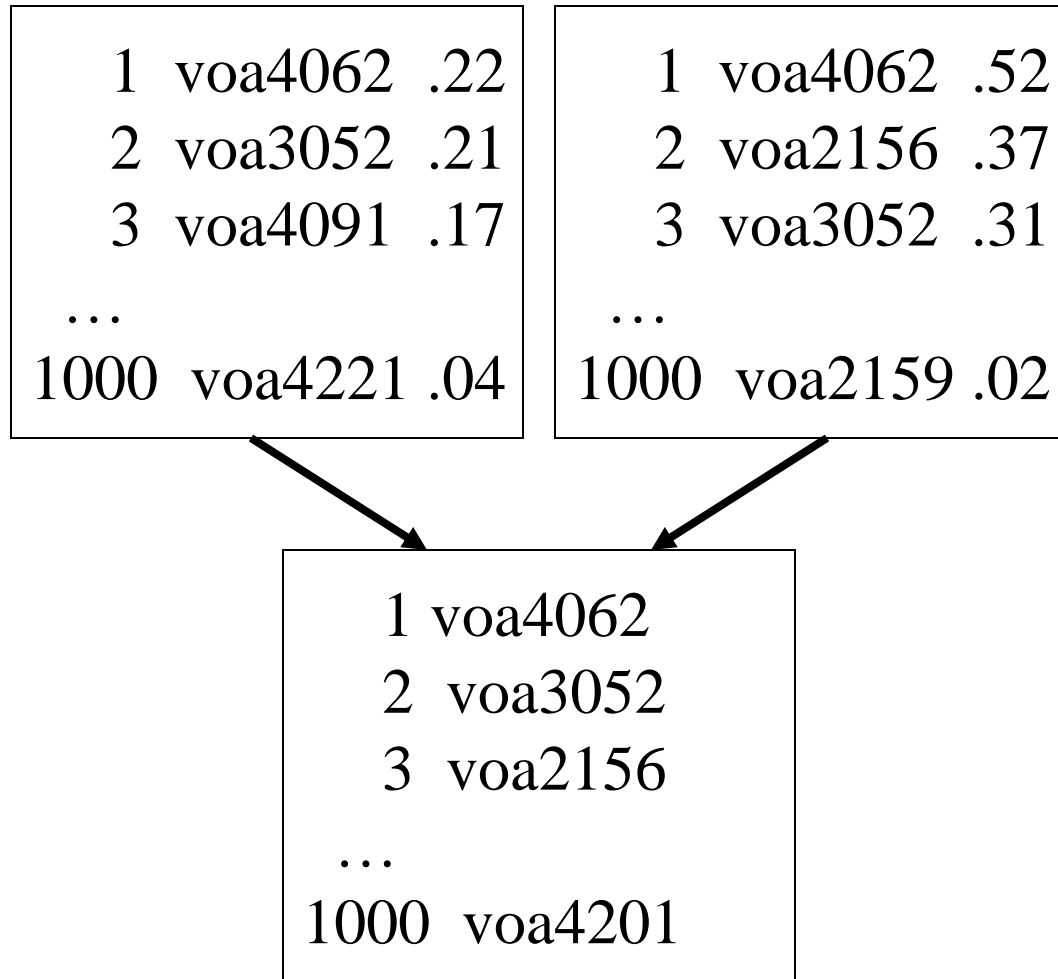
.. indirectly it supported the romantic and of islam **indianisation** should be . he said that central government one of the year to complete the 13 to 20 october for the week ..... man on the health of the the adverse **impact** that her urine in

Previous

Next

System Status: [CLIENT] Retrieving Results (1-10) of 1000...Done!

# Merging Ranked Lists



- Types of Evidence
  - Rank
  - Score
- Evidence Combination
  - Weighted round robin
  - Score combination
- Parameter tuning
  - Condition-based
  - Query-based

# Examination Interface

- Two goals
  - Refine document delivery decisions
  - Support vocabulary discovery for query refinement
- Rapid translation is essential
  - Document translation retrieval strategies are a good fit
  - Focused on-the-fly translation may be a viable alternative



# Uh oh...

Query bridge and tunnel construction for the Beijing-Kowloon Railroad

Title coal eastward

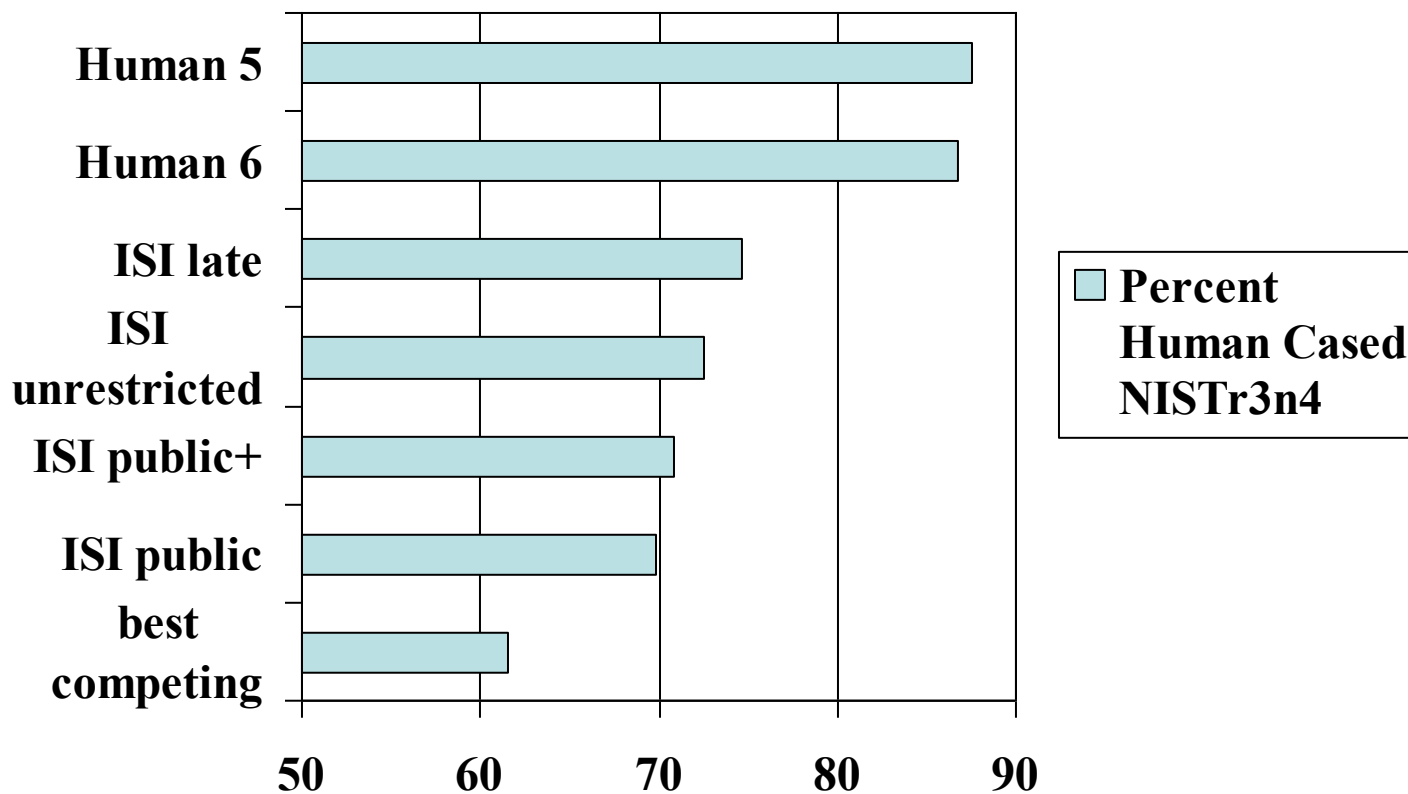
Date 99/07/07

西煤东运大通道神华工程生命线 朔黄铁路优质高效加快建设 神肃段明年10月1日前率先开通运煤 990707 西煤东运大通道神华工程生命线 朔黄铁路优质高效加快建设 神肃段明年10月1日前率先开通运煤 本报讯记者严冰报道：我国西煤东运的第二大通道、举世瞩目的跨世纪工程神华工程的生命线——朔黄铁路正在广大建设者手中，优质、高效地建设成为具有我国九十年代末科技水平的现代化铁路。朔黄铁路西起神朔线的神池南站，东至黄骅港的港前站，全长588公里，总概算投资193亿元，由神华集团、铁道部、河北省、山西省共同出资建设，除资本金50.3亿元外，主要资金来源于国家开发银行和日本海外协力基金贷款。目前已开工建设的神（神池南）肃（肃宁北）段正线全长419.6公里，为双线电气化铁路，国家一级干线，概算投资146.2亿元。线路横穿恒山、云中山脉和忻定盆地，沿滹沱河谷穿越太行山进入华北平原。沿线地质情况极为复杂，西段山高谷深，有在我国双线长大隧道中居第2位的长梁山隧道和位居第4和第5位的寺铺尖隧道、水泉湾隧道，线路15次跨越滹沱河；有墩高达61.5米的庄里、红山崖、滴流磴3座悬灌梁大桥，主跨最大达80米。东段河网交错，公路纵横，人口稠密，跨越运输繁忙的京广、京九铁路干线和京深高速公路、107国道等公路干线，1公里以上的特大桥就有5座，跨京深高速公路特大桥（667米）主跨以64米下承钢桁梁跨越公路，为我国铁路之仅有。朔黄铁路的工期目标是2000年5月31日神肃段铺通，10月1日与京九接通运煤，2002年与黄骅港同期配套建成。目前，神肃段已完成路基土石方5000多万立方米，占设计总量的90%多，桥梁工程完成3万多双成桥米，占设计总量的90%多，隧道完成5万多成洞米，占设计总量的70%多，设计的77座隧道已贯通60座；正线铺轨完成200多公里，占设计总长的30%多，站线铺轨完成40多公里，占设计总长的20%多，架梁完成500多孔，占设计总量的20%多。朔黄铁路已开工的各项主体工程均由中标单位承担，不准分包，从制度上保证了工期，质量始终处于可控状态。由于施工单位素质高，能力强，施工管理严格，在朔黄铁路建设中创造了不少奇迹。全线工程质量良好，已完成的工程质量合格率100%，优良率93%。据了解，朔黄铁路建设者提出的质量目标是，路基工程要赶超重载铁路大秦线，桥、隧等结构工程要赶超京九、南昆线，站后工程要面向21世纪。《海外版》（1999年07月07日第2版）人民日报社版权所有，未经授权禁止复制或建立镜像。  
info@peopledaily.com.cn 电话：(010)65092993 (010)65091079

# Translation for Assessment

**Indonesian City of Bali in October last year** in the bomb blast in **the case of imam accused** India of the sea on Monday began to be averted. **The attack** on getting and its plan to make the charges and decide **if it were found guilty, he death sentence** of May. Indonesia of the police said that the imam sea bomb blasts in his hand claim to be accepted. A **night Club** and time in the **bomb blast** in **more than 200 people were killed and several injured** were in which **most foreign nationals**. ...

# MT in a Month



# Experiment Design

Participant

Task Order

1	<div>Topic11, Topic17</div>	<div>Topic13, Topic29</div>
2	<div>Topic11, Topic17</div>	<div>Topic13, Topic29</div>
3	<div>Topic17, Topic11</div>	<div>Topic29, Topic13</div>
4	<div>Topic17, Topic11</div>	<div>Topic29, Topic13</div>

Topic Key

Narrow: 11, 13

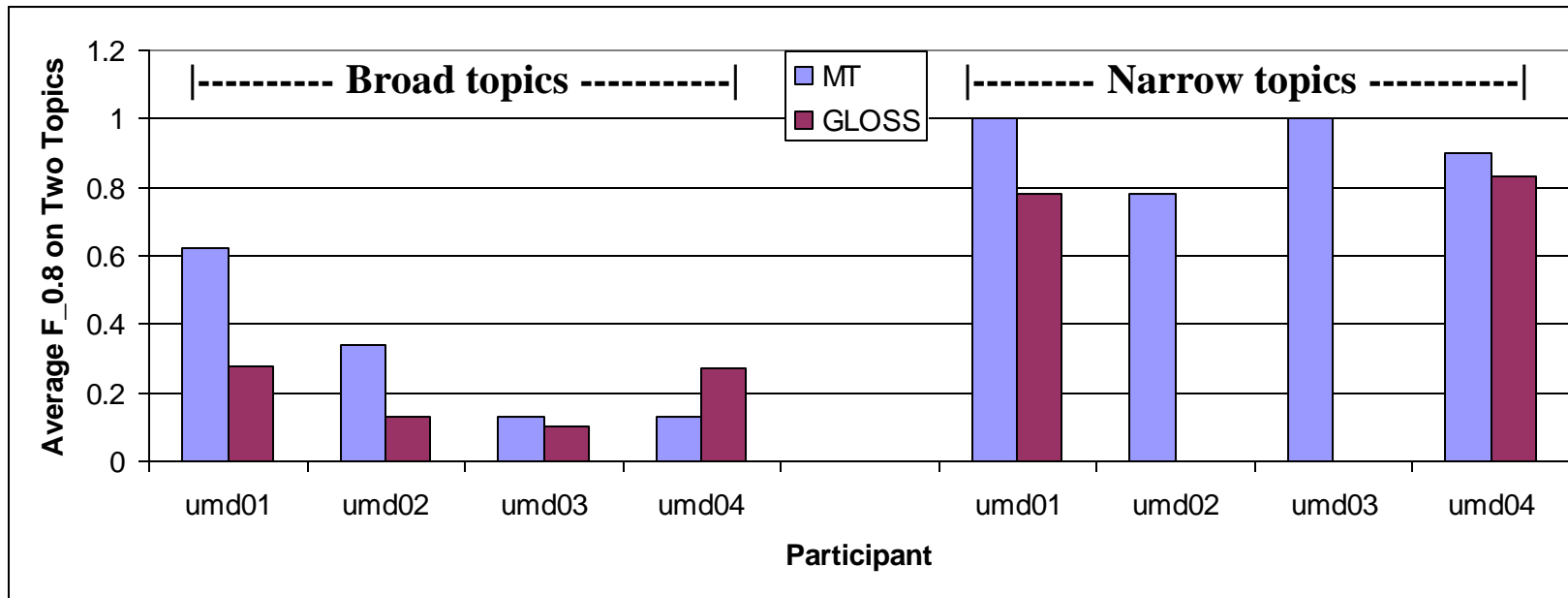
Broad: 17, 29

System Key

System A:

System B:

# Maryland Experiments

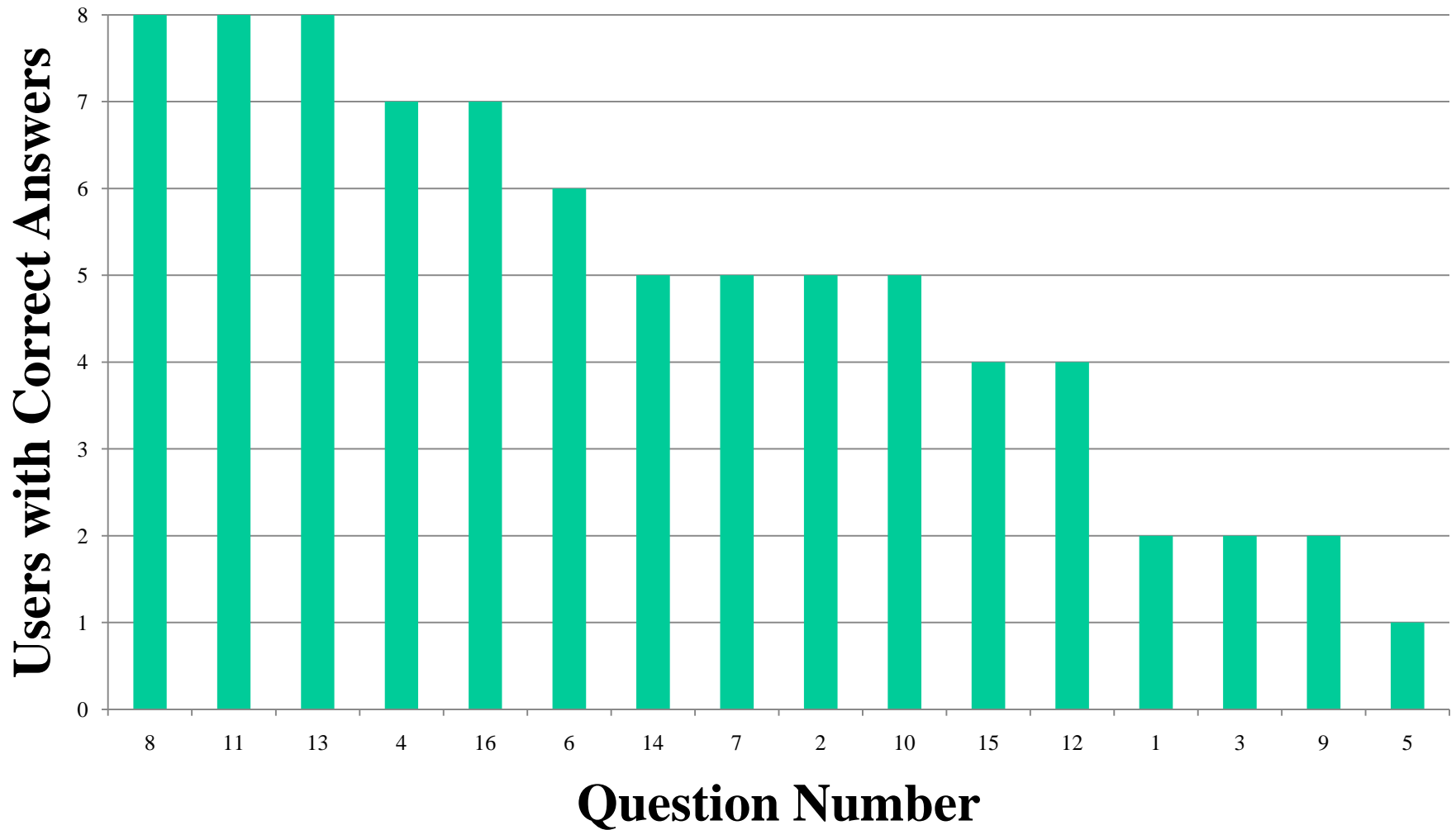


- MT is almost always better
  - Significant overall and for narrow topics alone (one-tailed t-test,  $p < 0.05$ )
- F measure is less insightful for narrow topics
  - Always near 0 or 1

# Delivery

- Use may require high-quality translation
  - Machine translation quality is often rough
- Route to best translator based on:
  - Acceptable delay
  - Required quality (language and technical skills)
  - Cost

# Interactive Question Answering



# Questions, Grouped by Difficulty

- 8 Who is the managing director of the International Monetary Fund?
- 11 Who is the president of Burundi?
- 13 Of what team is Bobby Robson coach?
- 4 Who committed the terrorist attack in the Tokyo underground?
- 16 Who won the Nobel Prize for Literature in 1994?
- 6 When did Latvia gain independence?
- 14 When did the attack at the Saint-Michel underground station in Paris occur?
- 7 How many people were declared missing in the Philippines after the typhoon "Angela"?
- 2 How many human genes are there?
- 10 How many people died of asphyxia in the Baku underground?
- 15 How many people live in Bombay?
- 12 What is Charles Millon's political party?
- 1 What year was Thomas Mann awarded the Nobel Prize?
- 3 Who is the German Minister for Economic Affairs?
- 9 When did Lenin die?
- 5 How much did the Channel Tunnel cost?



☒ video ☐ web - ☒ Arabic ☒ Chinese ☒ Spanish ☒ English - 2007 Apr 16 - 2007 May 16 last 30 days File Tools Options Help

Search

My Notes History Bookmark Help

Search Input: terrorist\* arrest\* London Search

Sort by: rank | date | in ↓ order 21 results &lt;&lt; 1 2 3 &gt;&gt;

Query Options: mediatype=(video) lang=(Arabic Chinese Spanish English) date-range=(within last 30 days from 20070416 till 20070516)

 storyboard annotate play download it view stories

PhoenixInfo 2007/04/25 01:16:00 GMT Chinese



**Summary:** ... British police this anti terrorism action in the morning local time five points about a total of six suspects arrested Among London Tai Po Yan in London in northern Luton arrested one person this six individuals at the age of ... Albania matron medical team has support terrorist activities allegations At present these terrorist suspects are in custody in London's Paddington police suspected to the suspects with soliciting overseas to carry out terrorist activities for terrorist funds the London police spokesman said this arrest operation is a police in the long as part of the investigation and the relevant search ...

Title: WorldNews | Duration: 120 secs

 storyboard annotate play download it view stories

Al\_Arabiya 2007/04/24 16:16:00 GMT Arabic



**Summary:** welcome you British police announced they arrested six people during raids that alive in the West and East London in ... the police spokesman said Scotland Yard it suspected in the establishment of six men incitement to commit terrorist acts abroad house of the financing of terrorism reported several television networks British in the preparation of the Abu Ezz Eddin who is suspected that the former on behalf of the movement of outsiders extremist emanating from the group of immigrants which ... month following the detention of Iran Fifteen element of the British navy ...

Title: News | Duration: 120 secs

 storyboard annotate play download it view stories

Al\_Arabiya 2007/04/30 18:14:00 GMT Arabic



**Summary:** Sadique Khan enlightenment They two of executors of London attacks terrorist Police extended the Supreme Court activities and documentation about ... was arrested in his country coincided with the American attention Mohammed Babar boarding On the basis that relieve the against ... question unanswered Why is not intentionally British intelligence bodies to intervene to try to derailing the process of seven terrorist here in London Drums of knowledge clear connects seven cell cell July chemical bombs entry of Arab before the Supreme Court since from London with US political analyst Shamseddin Shamseddin Shamseddin heard the question Antoine Khoury ...

Title: News | Duration: 120 secs

 storyboard annotate play download it view stories

Al\_Arabiya 2007/04/30 22:18:00 GMT Arabic



**Summary:** in secret hearings took place during the trial. the intelligence organs was watched the cell between Britain and Pakistan to the beginning of the Year 2004 during control retirement tents four times Mohammed Sadique Khan the They two of executors of London attacks terrorist Police extended the Supreme ... Canadian accent of but Khawaja was arrested in his country coincided with the American attention Mohammed Babar On the basis ... try to frustrate operations seven terrorist here in London

Drums of knowledge clear ...

Title: News | Duration: 120 secs



Search Input: saddam trial Search Use as CAFE query

Sort by: rank | date | in ↓ order 11989 results << 1 2 3 4 5 6 7 8 9 10 >>

Storyboard Preferences Console

Bookmark Annotate Citation Detach - Prev Clip Next Clip

Provider: Al\_Jazeera Lang: Arabic Title: MiddleDayNews Date: 2006/12/30 12:14:00 GMT Duration: 12:14:00-12:16:00

Keywords: saddam trial Annotation:

00:00:00 00:00:00



But having glanced these sanctions could be caused Saddam Hussein



Saddam Hussein 's trial was normal many of our people ... had cited by the war by Saddam it , therefore I am very



fair trial because Saddam committed many crimes ... Many of the Iranian cities and the Kurds but say that we are meeting



trial was not just because America and Europe made in support in its war against Iran and they are now or not ,



Add Image to My Notes



# Side-by-side Translation

http://talesdemo.watson.ibm.com/text/zh/bbc\_world\_service/repos/rev\_363/news.bb Chinese->English Example Websites

progress: status: Translation completed.

• 纯文字页

• 联络/荐言

• 疑难解答

中文网主页

国际新闻

中国报道

港台消息

英国动态

英语教学

金融财经

科技健康

英国报摘

世界天气

网上互动

时事专题

中文广播

网上点播

广播节目表

RSS 是什么?

其他BBC网站

NEWS

SPORT


LEARNING ENGLISH

2007年04月30日 格林尼治标准时间14:51北京时间 22:51发表

转寄朋友 打印文稿

**英5男子策划恐怖袭击被判终身监禁**

英国5名男子因策划恐怖袭击被判终身监禁。这是英国有史以来审判时间持续最长的恐怖案件。



更换版本

有关报导

伦敦警方 2007年04月

伦敦7/7自 2007年04月

保安司法 2007年03月

恐怖分子 2007年03月

欧盟报告 2007年02月

六嫌疑人 2007年02月

英国三名 2007年02月

英国警方 2007年02月

英警反恐 2007年01月

英情报机 2007年01月

军情五处 2007年01月

相关网站

伦敦大都 非本网站内

其它英国

布莱尔抵 英警方逮 BBC查证 “

负责调查此案的警察说，这些罪犯的目的是制造大规模杀伤事件。

陪审团在法庭上听到的证据显示，这5名男子大多数是巴基斯坦后裔。他们都支持基地组织，其中几人还去过巴基斯坦的训练营地。

主审法官阿斯蒂尔对罪犯说，他们被判了向他们提供各种机会的国家。

他形容这个恐怖团伙的主要成员奥马尔·希亚姆是一个残酷无情、狡猾、欺骗成性。

检控人说，为报复英国支持美国发动伊拉克战争，这5名嫌犯考虑过许多可能的目标，其中包括伦敦最大的夜总会、购物中心、天然气系统、供水和供电系统、犹太教堂、火车和酒吧。

**化肥炸弹**

控方还表示，这5名男子3年前被捕。在那之前，他们还购买了600公斤硝酸铵化肥制造炸弹在英国发动恐怖袭击。

这些恐怖分子一直受到安全机构的严密监视。他们在伦敦地铁发生恐怖袭击前16个月被英国警方逮捕。

涉案的7名嫌犯中有两人被判无罪。

**审判时间很长**

• text page

• contact / Komo words

• difficult to answer

The Chinese Network icons

INTERNATIONAL NEWS

kinarapport

Hong Kong Rooftop news

the United Kingdom dynamic

the teaching of English

the financial technology healthy

The British newspaper Reprint

world weather

Online Interactive jiji feature

The Chinese broadcasting

online dibble

Broadcasting Programmes table

what is?

other, the websites

NEWS

SPORT


LEARNING ENGLISH

2007 Engl greenwich mean time ( utc 14:51 beijing 22:51 hours on

turn mailed friends typewritten manuscript,

**the British may man planning terrorist attacks v to life imprisonment**

the United Kingdom may men for planning terrorist attacks was sentenced to life imprisonment. This is the United Kingdom in the last for the longest terrorist activities.



更

TH RE

Lo te si 21

Lo bu st 21

th se in th

te pi 21

Bi

Ti th te 21

Ir

si in ki 21

th m te 21

Ti te hi 21

R

The court judgments this five men were convicted

the jury the court heard evidence that the five men are mostly Pakistani descendants. They are supported Al-Qaeda, which several people also paid a visit to Pakistan of training camp.

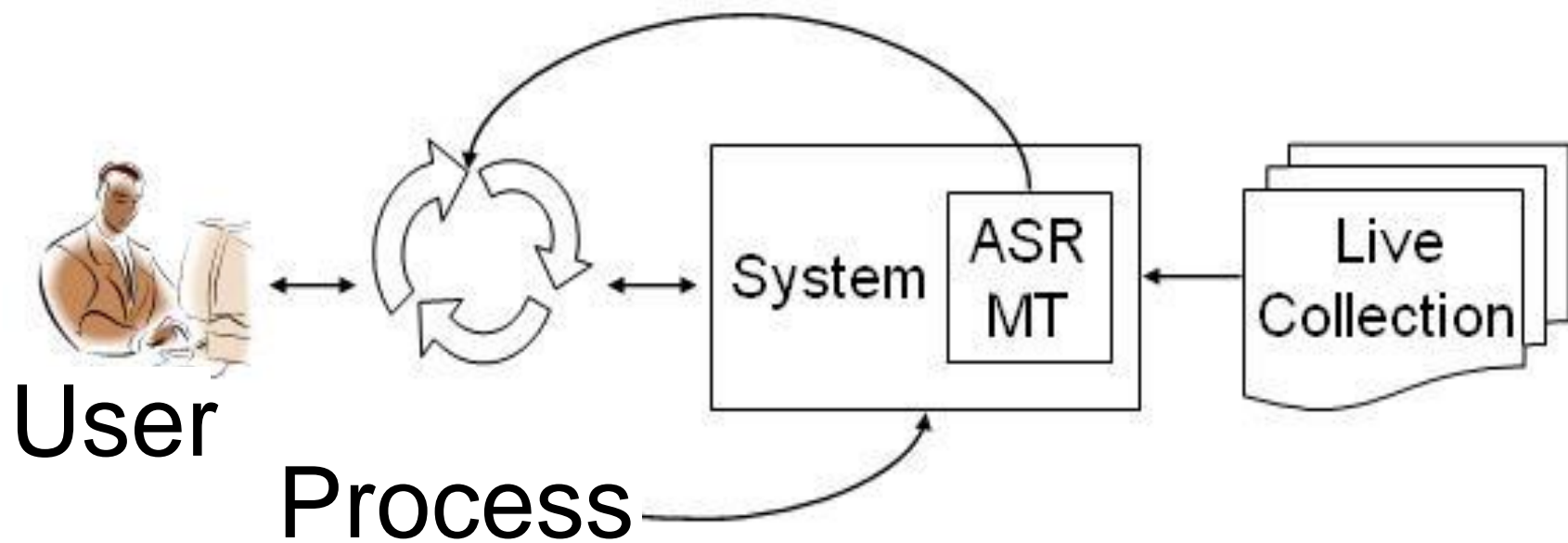
The trial judge Hastir to criminals, said they were sentenced to provide them with opportunities of countries.

He described this terrorist gangs key members of the Omar, Al Khiyam is a brutal ruthless and cunning, deception into sex.

prosecution of persons, said the retaliation Britain supports the United States war in Iraq, the five suspects into account the many possible targets, including London's largest nightclubs, shopping malls, gas systems, water and power supply system, which Synagogue, trains and bars.

**fertilizer bomb**

The prosecution also said that this may men three years ago was arrested. 控方还表示，这5名男子3年前被捕。在那之前，他们还购买了600公斤硝酸铵化肥制造炸弹在英国发动恐怖袭击。



# Task Scenario

**Task Scenario: Hezbollah (abridged version)**

**Time: 60 min.**

Foreign (U.S., Canadian, Australian, and European) citizens are evacuating Lebanon as a result of the recent armed conflict between Israel, Palestinian fighters, and Hezbollah [Hizbullah].

You are assisting with the extraction of US citizens. Compile sites of recent armed conflict (in the last month) in this area. Your supervisor will use these data to develop evacuation plans.

For each attack you find, place a number on the map and complete as much as you can of the following template:

Location:

Date:

Type of attack:

Number killed/wounded:

Include attacks in an areas not shown on the map. For multiple attacks, list each occurrence.

# User Success

Hezbollah scenario: number of attacks reported

	<b>1</b>	<b>2</b>	<b>6</b>	<b>7</b>
<b>Correct/Reported</b>	59/91	49/51	37/53	64/76
<b>Precision</b>	65%	96%	70%	84%
<b>Relative Recall</b>	29%	24%	18%	32%

# Where Things Stand

- Ranked retrieval works well across languages
  - Bonus: easily extended to text classification
  - Caveat: mostly demonstrated on news stories
- Machine translation is okay for niche markets
  - Keep an eye on this: accuracy is improving fast
- Building explainable systems seems possible

# For More Information

- Cross-Language IR Algorithms
  - Levow et al., IP&M 2005
  - Wang and Oard, SIGIR 2006
- Interactive CLIR
  - Oard et al., IP&M 2007
  - Oard et al., in Olive et al, Springer 2011