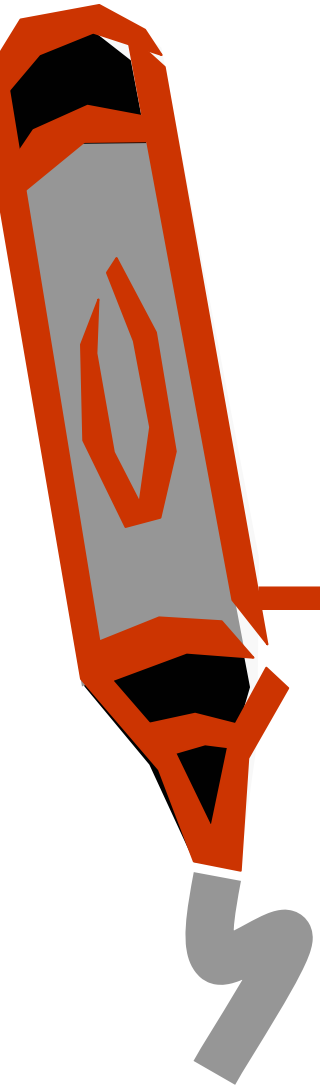


Cross-Language IR

many slides courtesy
James Allan@umass
Jimmy Lin, University of Maryland
Paul Clough and Mark Stevenson, University of Sheffield, UK

What is Cross-Lingual Retrieval?

- Accepting questions in one language (English) and retrieving information in a variety of other languages
  - "questions" may be typical Web queries or full questions in across-lingual question answering (QA) system
  - "information" could be news articles, text fragments orpassages, factual answers, audio broadcasts, written documents, images, etc.
- Searching distributed, unstructured, heterogeneous, multilingual data
- Often combined with summarization, translation, and discovery technology

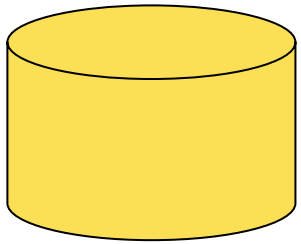
Current Approaches to CLIR

- Typical approach is to translate query, use monolingual search engines, then combine answers
  - other approaches use machine translation of documents
  - Or translation into an interlingua
- Translation ambiguity a major issue
  - multiple translations for each word
  - query expansion often used as part of solution
  - translation probabilities required for some approaches
- Requires significant language resources
  - bilingual dictionaries
  - parallel corpora
  - “comparable” corpora
  - MT systems

Two Approaches

- Query translation
  - Translate English query into Chinese query
  - Search Chinese document collection
  - Translate retrieved results back into English
- Document translation
  - Translate entire document collection into English
  - Search collection in English
- Translate both?

Query Translation



Chinese Document Collection

Retrieval Engine

Chinese documents

Chinese queries

Translation System

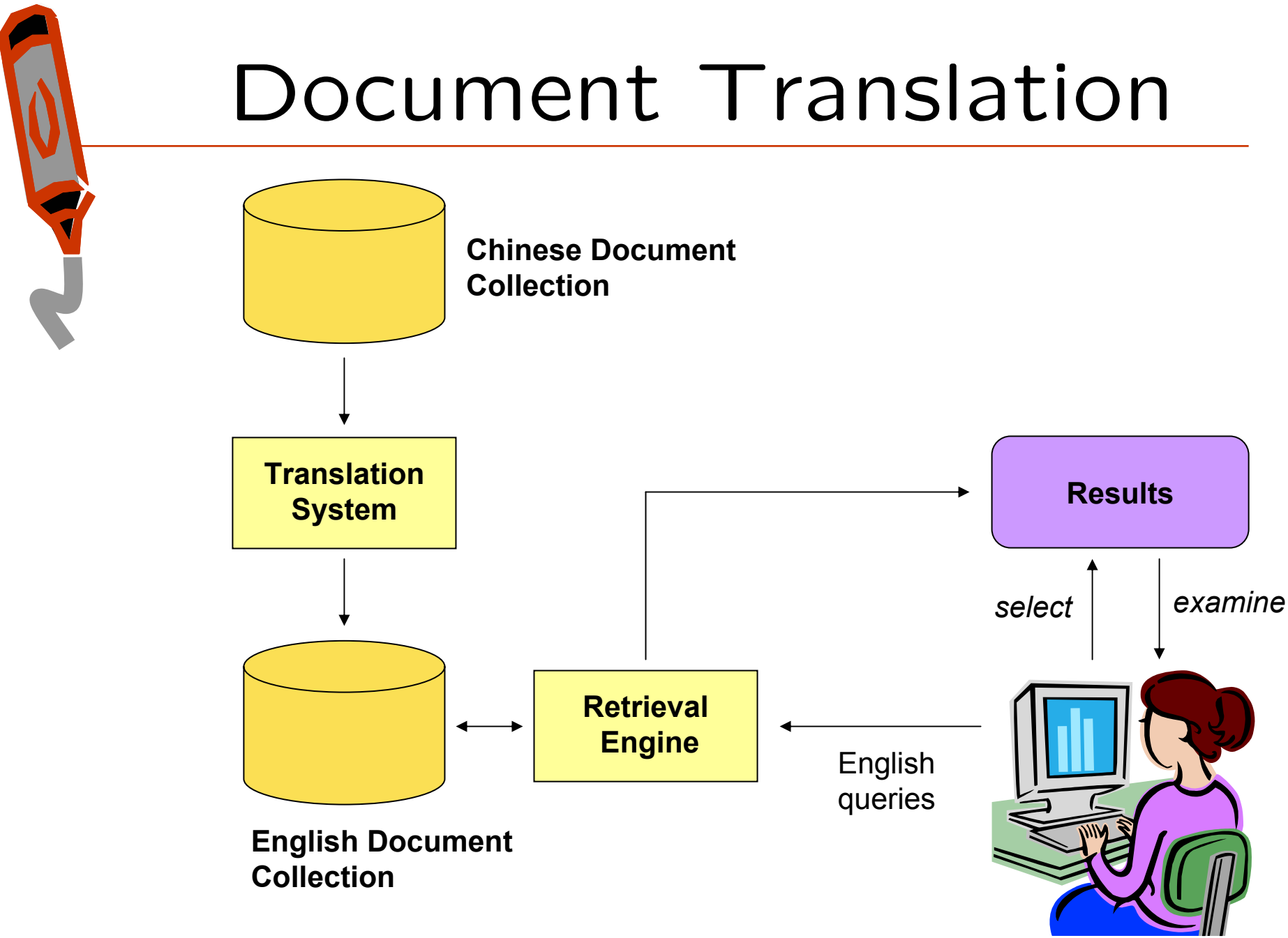
Results

select

examine

English queries

Document Translation



Tradeoffs

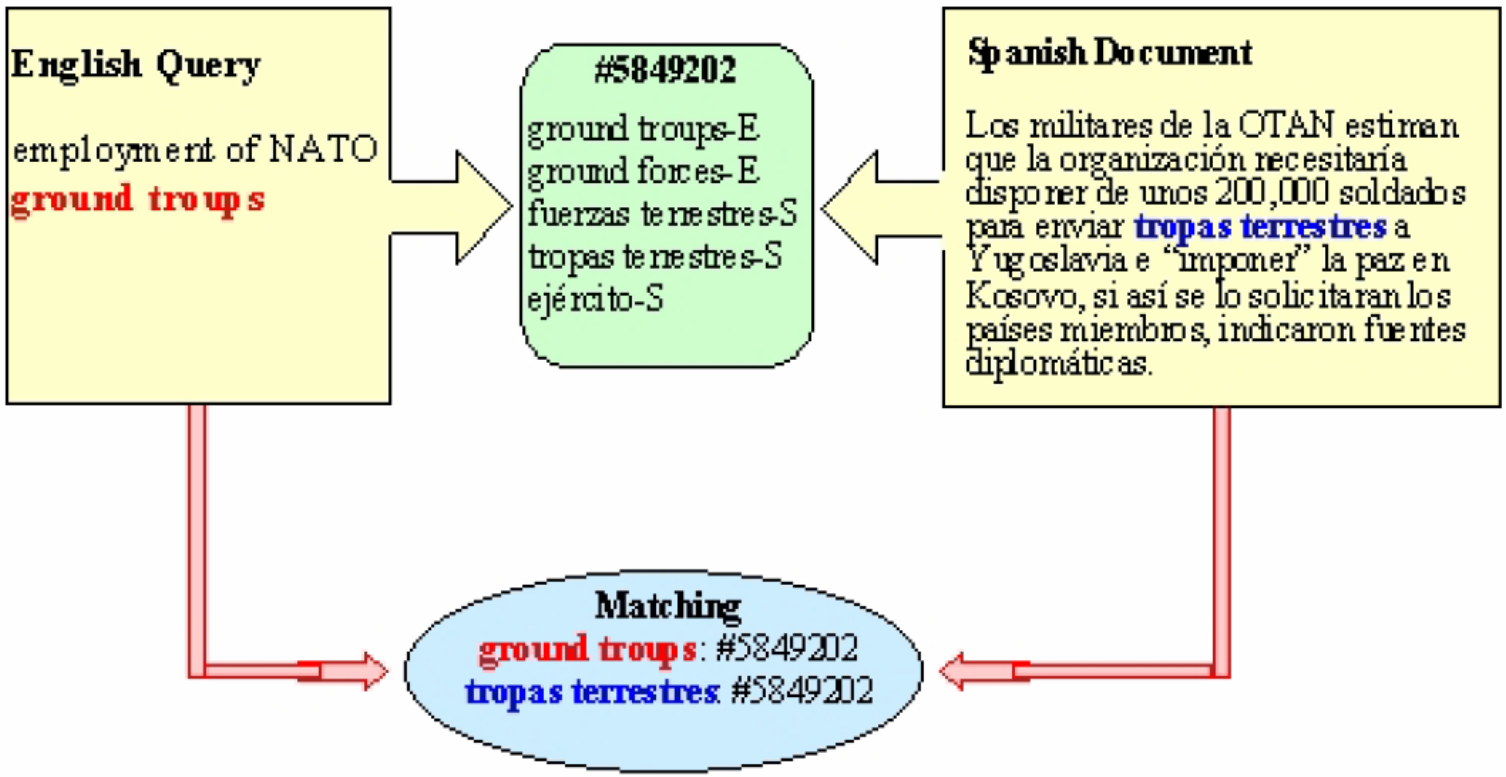
- Query Translation
  - Often easier
  - Disambiguation of query terms may be difficult with short queries
  - Translation of documents must be performed at query time
- Document Translation
  - Documents can be translate and stored offline
  - Automatic translation can be slow
- Which is better?
  - Often depends on the availability of language-specific resources (e.g., morphological analyzers)
  - Both approaches present challenges for interaction

A non-statistical approach

- A non-statistical approach
- Interlingua approaches
  - Translate query into special language
  - Translate all documents into same language
  - Compare directly
  - Cross-language retrieval becomes monolingual retrieval
- Choice of interlingua?
  - Could use an existing language (e.g., English)
  - Create own
- Textwise created a “conceptual interlingua”

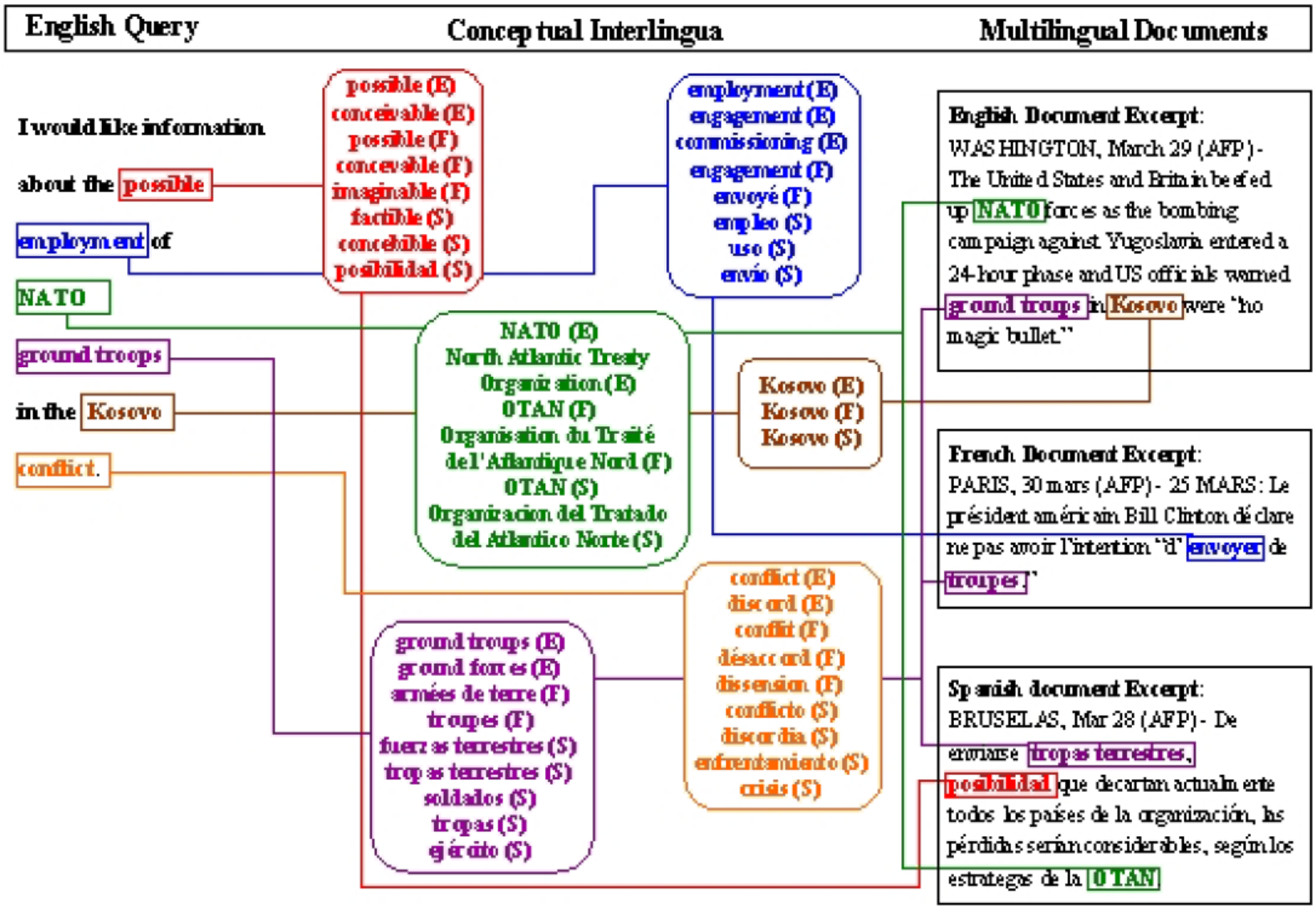
CINDOR

- Conceptual Interlingua for Document Retrieval



[Liddy, Infonortics 1999]

CINDOR



[Liddy, Infonortics 1999]

Does it work?

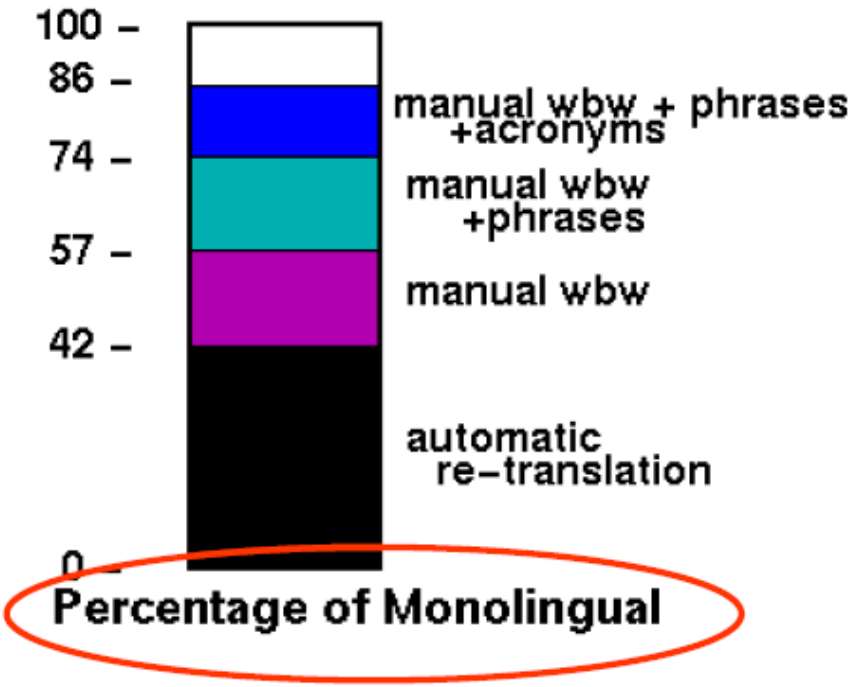
- Some background research suggested large gains over word-by-word translation
- Fielded in TREC-7 cross-language task
- Performed poorly overall
  - System not completed at the time
  - Interlingua incomplete
  - Several small processing errors added up
  - On queries without problems, comparable to monolingual
- Statistical methods now dominate the field

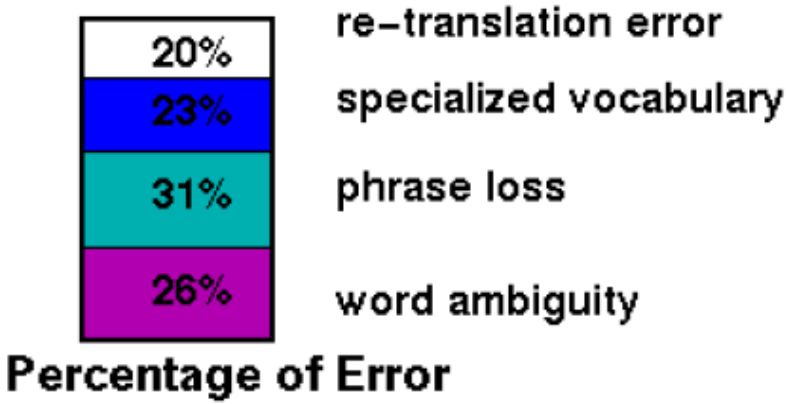
Current Capabilities of CLIR

- Best performance obtained by
  - probabilistic approach using translation probabilities estimated from an aligned parallel corpus
  - "structured" query that treats translations from bilingual dictionary as synonyms and uses advanced search engine
  - Combination of techniques including MT
  - Most experiments done in Chinese, Spanish, French, German, and recently, Arabic

- Cross-lingual can achieve 80-90% effectiveness of monolingual
  - with sufficient language resources
  - sometimes does even better, but can also do worse

CLIR errors


100 –
86 –
74 –
57 –
42 –
0
manual wbw + phrases +acronyms
manual wbw +phrases
manual wbw
automatic re–translation
Percentage of Monolingual


20%
23%
31%
26%
re–translation error
specialized vocabulary
phrase loss
word ambiguity
Percentage of Error

But how good is “monolingual”?

- Not easy to summarize IR performance as a single number
  - We’ve considered average precision, Swet’s number, utility functions, expected search length, ...
- Based on measures of recall and precision...
  - Breakeven of 30% for “Web” queries, precision 40% in top 20, 20% in top 100
  - Breakeven of 45% for “analyst” queries, precision 65% in top 20, 45% in top 100
  - Recall can be improved through techniques such as query expansion and relevance feedback

Adding New Languages

- Morphological processing
  - segmenting (what is a word?)
  - stemming (combining inflections and variants)
  - stopwords (words that can be ignored)
- Language resources
  - minimum is a bilingual dictionary
  - parallel or comparable corpora are even better
  - MT system is a luxury

Problems with CLIR

- Morphological processing difficult for some languages (e.g. Arabic)
  - Many different encodings for Arabic
    - Windows Arabic (e.g. dictionaries)
    - Unicode (UTF-8) (e.g. corpus)
    - Macintosh Arabic (e.g. queries)
  - Normalization
    - Remove diacritics

      العَرَبِيَّة to العربية *Arabic (language)*
  - Standardize spellings for foreign names

    كلنتون vs كلينتون *"Kleentoon" vs "Klntoon" for Clinton*

Problems with CLIR

- Morphological processing (contd.)
  - Arabic stemming
  - Root + patterns+suffixes+prefixes=word

  ktb+CiCaC=kitab

- All verbs and nouns derived from fewer than 2000 roots

- Roots too abstract for information retrieval

  ktb → kitab *a book* kitabi *my book*
  alkitab *the book* kitabuki *your book (f)*
  kataba *to write* kitabuka *your book (m)*
  maktab *office* kitabuhu *his book*
  maktaba *library, bookstore ...*
  Want stem=root+pattern+derivational affixes?

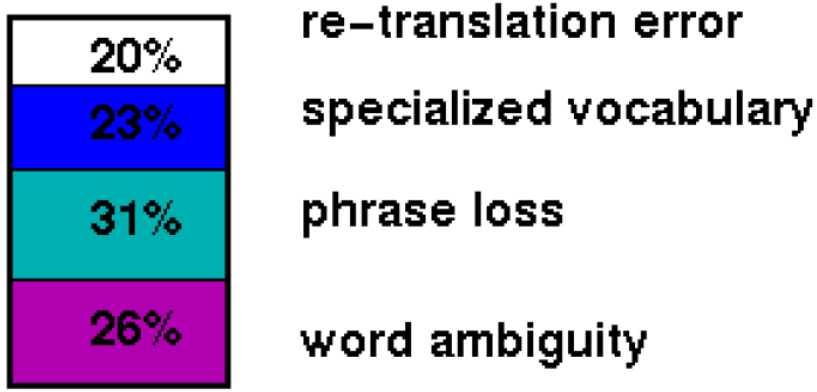
- No standard stemmers available, only morphological (root) analyzers

Problems with CLIR

- Availability of resources
  - Names and phrases are very important, most lexicons do not have good coverage
- Difficult to get hold of bilingual dictionaries
  - can sometimes be found on the Web
- e.g. for recent Arabic cross-lingual evaluation we used 3 on-line Arabic- English dictionaries (including harvesting) and a small lexicon of country and city names
  - Parallel corpora are more difficult and require more formal arrangements

Phrase translation

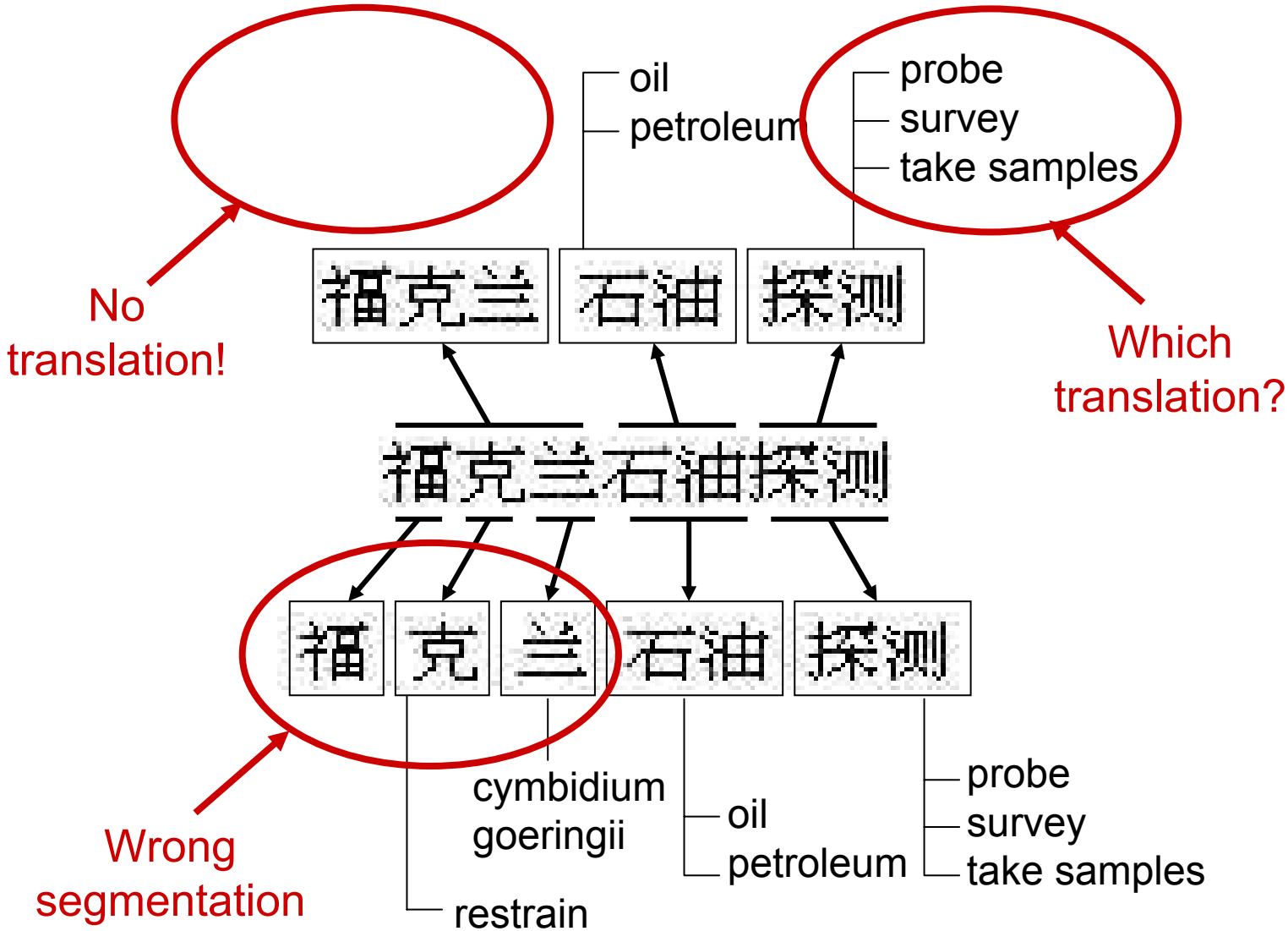
- Phrases are a major source of translation error
- How to get phrases translated properly?
- Assume that correct translations of words in phrase co-occur
  - Given two-word phrase “A B”
  - Look at all translations of A: A1 or A2 or ... or An (and B, similarly)
  - Look at all pairs “Ai Bj” and see which of them co-occur
- Probably in passages of the collection
  - Use the best pair as the phrase translation



Example Phrase

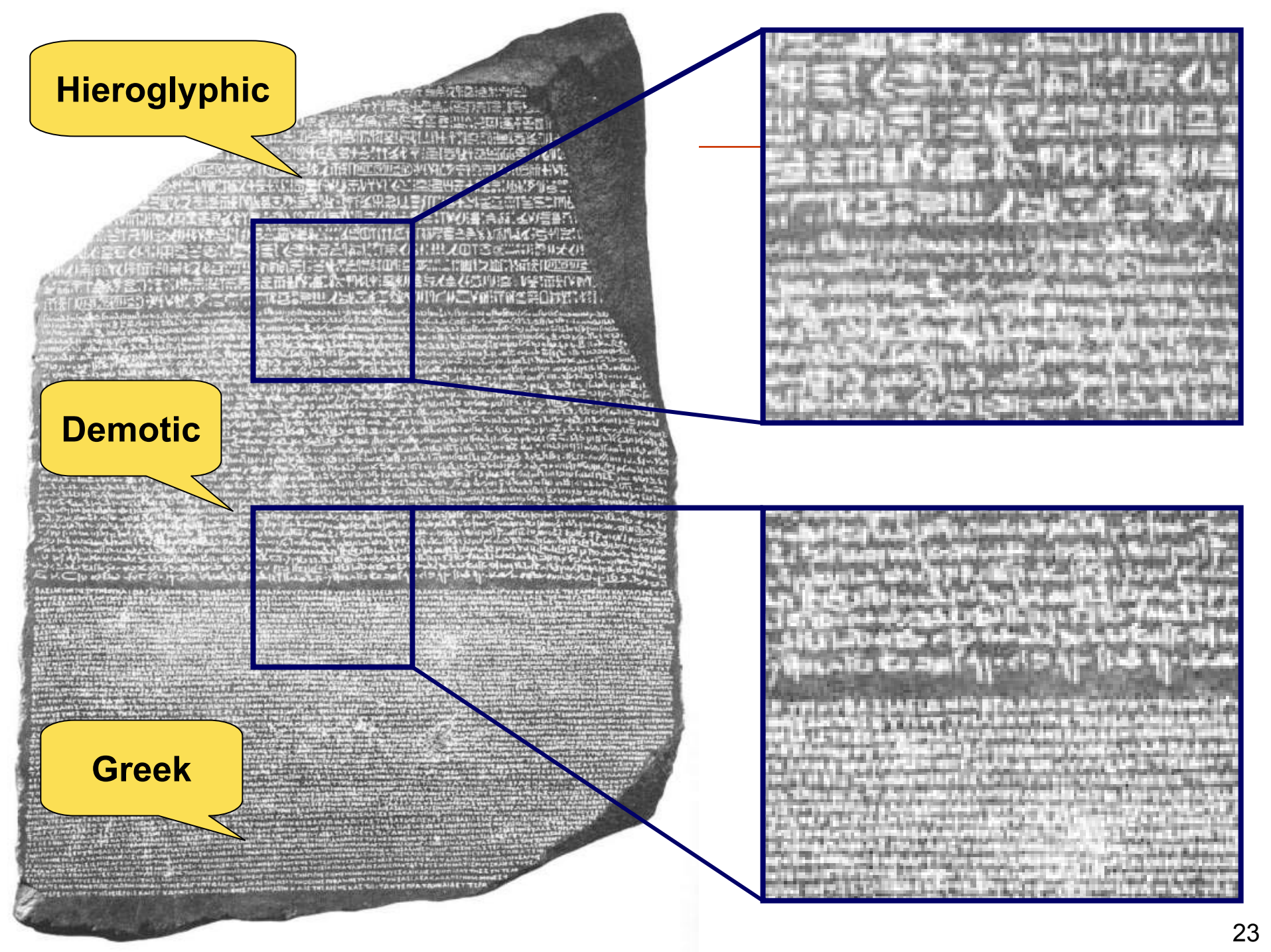
- Worked quite well in English-Spanish CLIR
- Consider Spanish phrase “Proceso Paz”
  - process, lapse of time, trial, prosecution, action, lawsuit, proceedings, processing
  - peace, peacefulness, tranquility, peace, peace treaty, kiss of peace, sign of peace
- Ranked possible translation pairs:
  - peace process
  - peacefulness process
  - tranquility process
  - ...

CLIR Issues


No translation!
oil
petroleum
probe
survey
take samples
Which translation?
福克兰
石油
探测
福克兰石油探测
福
克
兰
石油
探测
Wrong segmentation
cymbidium goeringii
restrain
oil
petroleum
probe
survey
take samples

Learning to Translate

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


Hieroglyphic
Demotic
Greek

Word-Level Alignment

English

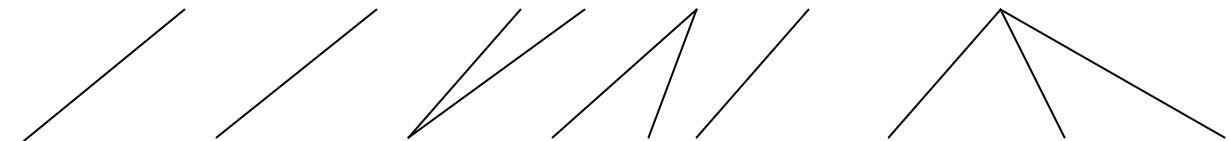
Diverging opinions about planned tax reform

Unterschiedliche Meinungen zur geplanten Steuerreform

German

English

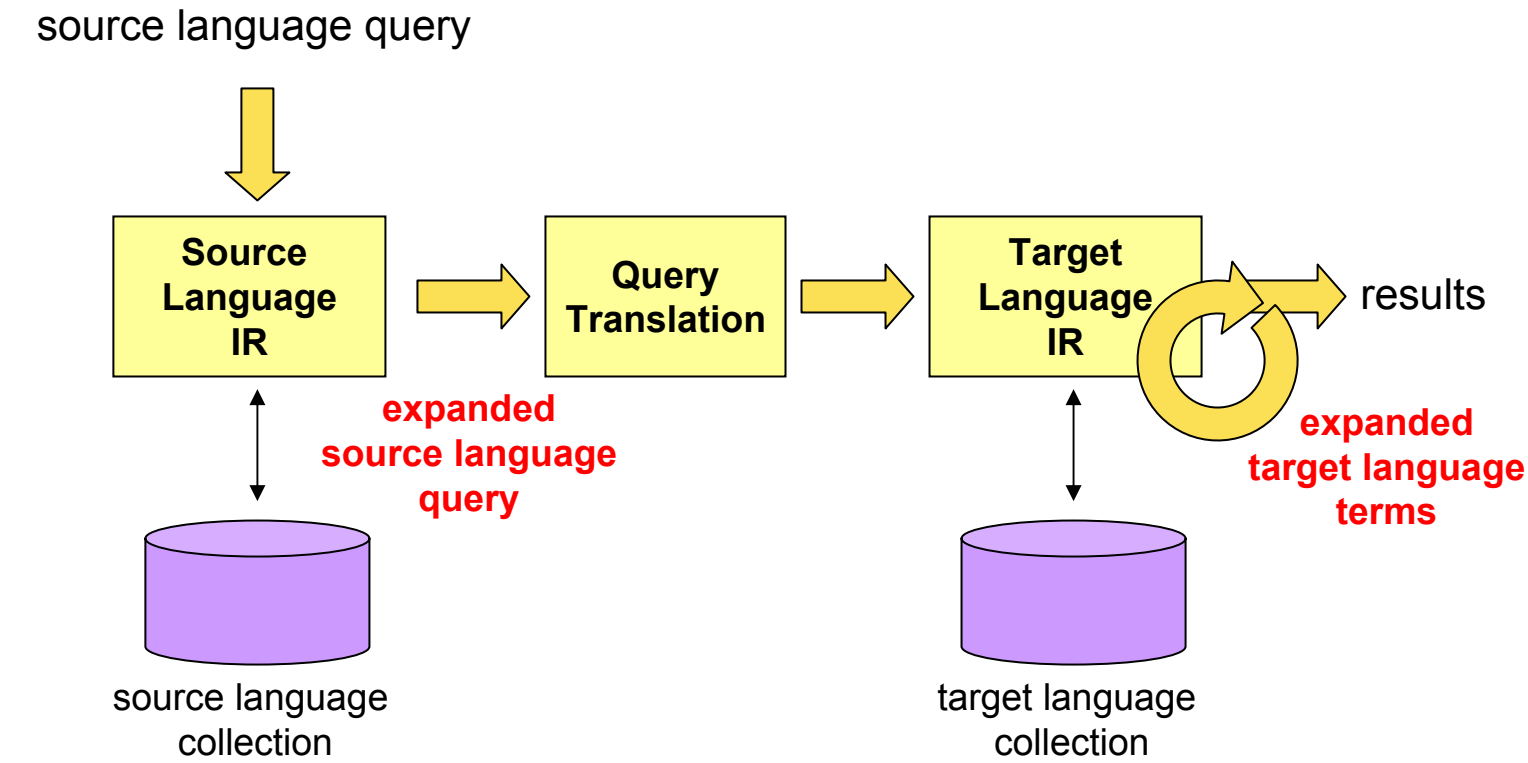
Madam President , I had asked the administration …



Señora Presidenta, había pedido a la administración del Parlamento …

Spanish

Query Expansion/Translation



TREC 2002 CLIR/Arabic

- Most recent (US-based) study in CLIR occurred at TREC
  - Results reported November 2002
- Problem was to retrieve Arabic documents in response to English queries
  - Translated Arabic queries provided for monolingual comparison
- Corpus of Arabic documents
  - 896Mb of news from Agence France Presse
  - May 13, 1994 through Decmber 20, 2000
  - 383,872 articles
- Topics
  - 50 TREC topic statements in English
  - Average of 118.2 relevant docs/topic (min 3, max 523)
- Nine sites participated
  - 23 CL runs, 18 monolingual

Sample topic

<top>
<num>Number: AR26</num>
<title>Kurdistan Independence</title>
<desc> Description:
How does the National Council of Resistance relate to the potential independence of Kurdistan?
</desc>
<narr> Narrative:
Articles reporting activities of the National Council of Resistance are considered on topic. Articles discussing Ocalan's leadership within the context of the Kurdish efforts toward independence are also considered on topic.
</narr>
</top>

sample topic arabic document

```
<DOC>
  <DOCNO>20000321_AFP_ARB.0001</DOCNO>
  <HEADER>ا0100 4 ش 8920 قبر /اف-ذزز03 اسرائيل/فلسطينيون</HEADER>
- <BODY>
    <HEADLINE>جرح ثلاثه اسرائيليين اصابة اثنين منهم خطيرة في هجوم في الضفة الغربيه</HEADLINE>
  - <TEXT>
      <P>القدس 12-3 (اف ب) - افادت حصيلة جديدة للجيش الاسرائيلي ان ثلاثة اسرائيليين جرحوا مساء امس الاثنين في هجوم جرى عندما اطلق
        عليهم الرصاص من سيارة تجاوزت السيارة المدنية التي كانت تقلهم قرب ترقوميه في محيط الخليل بالضفة الغربية.</P>
      <P>واوضح المتحدث باسم الجيش الاسرائيلي ان سائق السيارة التي كانت تقل الاسرائيليين، وهو من مستوطني الضفة الغربية اصيب بجروح
        "طفيفة". ووصف حالة احد الجريحين الاخرين بانها "حرجة" وحالة الثاني بانها "خطيرة".</P>
      <P>وتشكل الخليل حيث يقيم 004 مستوطن يهودي بحماية الجيش الاسرائيلي وسط 021 الف فلسطيني، بؤرة توتر بين الاسرائيليين والعرب. وقد
        انسحبت اسرائيل في كانون الثاني/يناير 7991 من 08% من هذه المدينة وابقت على وجود عسكري كبير في الحي الذي يسكنه المستوطنون
        .</P>
      <P>وجرح الاسرائيليون الثلاثة عندما تعرضت السيارة التي كانوا فيها لاطلاق نار من سيارة اخرى تجاوزتها قرب بلدة ترقوميه التي يؤدي اليها "الممر
        الامن" الذي يربط بين غزة وجنوب الضفة الغربية مرورا بالاراضي الاسرائيلية.</P>
      <P>وقد نقل الجريحان بسيارة اسعاف تم بمروحية الى مستشفى حداسا في القدس.</P>
      <P>وبدأ الجيش عمليات بحث عن الفاعلين واقام حواجز على الطرقات.</P>
      <P>وابلغت السلطة الفلسطينية بملابسات الهجوم لتحاول العثور على مرتكبيه.</P>
      <P>واشاد المسؤولو الاسرائيليون في الفترة الاخيرة بالتعاون مع اجهزة الامن الفلسطينية في اطار مكافحة الارهاب.</P>
      <P>ونشرت بلدية مستوطنه كريات اربع القريبة من الخليل بيان احتجاج على سياسة السلام التي يتبعها رئيس الوزراء الاسرائيلي ايهود باراك الذي
        اتهمه "بترك المستوطنين رهائن بايدي الفلسطينيين".</P>
      <P>وقال الجيش الاسرائيلي في تقديرات اولية ان خلية تابعة لحركة المقاومة الاسلامية (حماس) قد تكون وراء الاعتداء.</P>
      <P>وتعارض حركة حماس بشدة اتفاقات اوسلو حول الحكم الذاتي الفلسطيني المبرمة عام 3991 وقد اعلنت مسؤوليتها عن غالبية الاعتداءات
        التي استهدفت اسرائيل منذ ذلك الحين.</P>
    </TEXT>
    <FOOTER>شف/ا موا004 اف _____</FOOTER>
  </BODY>
  <TRAILER>405012 00 جمت مار</TRAILER>
</DOC>
```

Stemming

- TREC organizers provided standard resources
- To allow comparisons of algorithms vs. resources
- One of those was an Arabic stemmer
  - UMass developed a “light stemmer” also used heavily

Stemmer	Prefixes	Suffixes
Al-Stem	wAl, fAl, bAl, bt, yt, lt, mt, wt, st, nt, bm, lm, wm, km, fm, Al, ll, wy, ly, sy, fy, wA, fA, lA, and bA. (و ال ف ال ب ال ب ت ي ت ل ت م ت و ت س ت ن ت ب م ل م و م ك م ف م ال ل ل و ي ل ي ف ي و ا ف ا ل ا ب ا)	At, wA, wn, wh, An, ty, th, tm, km, hm, hn, hA, yp, tk, nA, yn, yh, p, h, y, A. (ا ت ، و ا ، و ن ، و ه ، ا ن ، ت ي ، ت ه ، ت م ، ك م ، ه م ، ه ن ، ه ا ، ي ة ، ت ك ، ن ا ، ي ن ، ي ه ، ة ، ه ، ي ، ا)
Umass Stemmer	Al, wAl, bAl, kAl, fAl, and w (ا ل و ا ل ب ا ل ك ا ل ف ا ل و)	hA, An, At, wn, yn, yh, yp, p, h, and y (ي ، ه ، ة ، ي ة ، ي ه ، ي ن ، و ن ، ا ت ، ا ن ، ه ا)
Modified UMass Stemmer	-- Identical to U Mass, plus ll, and wll (ل ل ، و ل ل)	-- Identical to U Mass --

Darwish and Oard, TREC 2002]

Stemming (Berkeley)

- Alternative way to build stem classes
- Trying to deal with complex morphology
- Use MT system to translate Arabic words
  - Now have (arabic, english) pairs
- Stop and stem all of the English words/phrases using favorite stemmer
  - (arabic, english-stem) pairs
  - If English stem is the same, then assume Arabic words should be in the same stem class
- (Also used a light stemmer)

Stemming

Arabic word	English translation	Arabic word	English translation	Arabic word	English translation	Arabic word	English translation
أطفال	children	اطفالهن	their children	بطفل	by child	فالطفلة	then the child
أطفالا	children	اطفالي	my children	بطفلة	by child	فطفل	then child
أطفالنا	our children	الأطفال	children	بطفلتنا	by our child	كأطفال	as children
أطفاله	and his children	الاطفال	children	بطفلته	by his child	كالطفل	as the child
أطفاله	his children	الطفل	the child	بطفله	by his child	لأطفال	to children
أطفالها	her children	الطفلان	the children	بطفلها	by her child	لطفلها	to her child
أطفالهم	their children	الطفلة	the child	بطفلهما	by their child	للطفلة	to the child
أطفالهن	their children	الطفلتان	the children	بطفلين	by children	وأطفالنا	and our children
أطفالي	my children	الطفلتين	the children	بطفليها	by her children	والأطفال	and the children
اطفال	children	الطفله	the child	طفل	child	وبطفل	and by child
اطفالا	children	الطفلين	the children	طفلا	child	وبطفلين	and by children
اطفالك	your children	بأطفال	by children	طفلان	children	وطفلة	and child
اطفالكم	your children	بأطفاله	by his children	طفلاها	her children	وطفلتان	and children
اطفالكن	your children	بأطفالها	by her children	طفلة	child	وطفلتا	and our child
اطفالنا	our children	بالأطفال	by the children	طفلت	child	وطفلها	and her child
اطفاله	his children	بالطفل	by the child	طفلتان	children	وطفليه	and his children
اطفالها	her children	بالطفلة	by the child	طفلته	his child	وطفليها	and her children
اطفالهم	their children	بالطفلتين	by the children	طفلتنا	our child	ولأطفالها	and to her children
اطفالهما	their children	بالطفلين	by the children	طفله	his child	وللطفل	and to the child

UMass core approaches

- InQuery
  - For each English word, look up all translations in dictionary
- If not found as is, try its stem
  - Stem all Arabic translations
  - Apply operators
- Put Arabic phrases in #filreq() operator
- Use synonym operator, #syn(), for alternate translations
- Wrap all together in #wsum() operator
- Cross-language language modeling (after BBN)

$$P(Q_e|D_a) = \prod_{e \in Q_e} \left(\alpha \sum_{a \in Arabic} P(a|D_a)P(e|a) + (1-\alpha)P(e|GE) \right)$$

Breaking the LM approach apart

- Query likelihood model
- $P(a|D_a)$
  - Probability of Arabic word in the Arabic document
- $P(e|a)$
  - Translation probability (prob. of English word for Arabic word)
- $P(e|GE)$
  - Smoothing of the probabilities

$$P(Q_e|D_a) = \prod_{e \in Q_e} \left(\alpha \sum_{a \in Arabic} P(a|D_a)P(e|a) + (1-\alpha)P(e|GE) \right)$$

Calculating translation probabilities

- Dictionary or lexicon
  - Assume equal probabilities for all translations
  - Unless dictionary gives usage hints
- Parallel corpus
  - Assume sentence-aligned parallel corpora

- Know that sentence E is a translation of sentence A
  - Estimate P(e|a) from those aligned sentences
  - Consider sentence pairs (E,A) where e is in E and a is in A
  - To get P(e|a), divide by number of Arabic sentences containing a

$$P(e|a) = \frac{|\{(E,A)|e \in E \text{ and } a \in A\}|}{|\{A|a \in A\}|}$$

Other techniques

- Query expansion
  - Useful to bring in additional related words
  - Same as in monolingual retrieval
- Expand query in English
  - Need comparable corpus (why comparable?)
  - Brings in synonyms and other words related to query
- Expand translated query in Arabic
  - Done on actual target corpus
  - Brings in Arabic synonyms not in dictionary
  - BBN in TREC 2002 was careful to expand only by translation of original query words
- Can do neither, either, or both
- UMass added 5 terms from English and 50 from Arabic
  - For LM runs, used “relevance modeling” instead in Arabic

CLIR better than IR?

- How can cross-language beat within-language?
  - We *know* there are translation errors
  - Surely those errors should *hurt* performance
- Hypothesis is that translation process may disambiguate some query terms
  - Words that are ambiguous in Arabic may not be ambiguous in English
  - Expansion during translation from English to Arabic prevents the ambiguity from re-appearing
- Has been proposed that CLIR is a model for IR
  - Translate query into one language and then back to original
  - Given hypothesis, should have an improved query
  - Should be reasonable to do this across many different languages

International Research Programs

- Major ones are
  - TREC(US DARPA under TIDES program),
  - CLEF(EU) and
  - NTCIR (Japan)
- Programs were initially designed for ad-hoc cross-language text retrieval, then extended to multi-lingual, multimedia, domain specific and other dimensions.

CL image retrieval, CLEF 2003

- A pilot experiment in CLEF 2003
- Called ImageCLEF
- Combination of image retrieval and CLIR
- An ad hoc retrieval task
- 4 entries
  - NTU (Taiwan)
  - Daedalus (Spain)
  - Surrey (UK)
  - Sheffield (UK)

Why a new CLEF task?

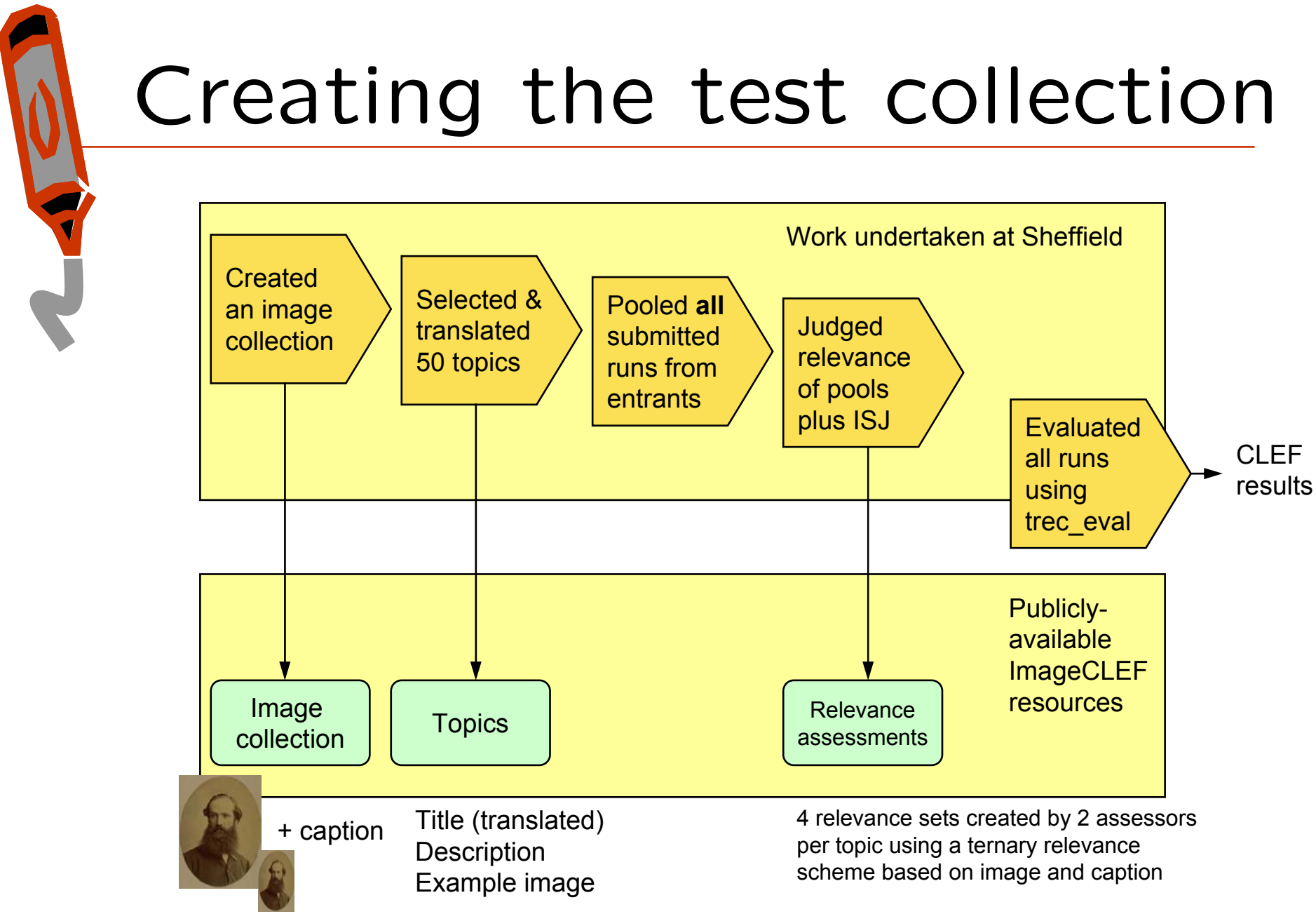
- No existing TREC-style test collection
- Broadens the CLEF range of CLIR tasks
- Facilitates CL image retrieval research
- International forum for discussion

CL image retrieval, CLEF 2003

Given a user need expressed in a language different from the document collection, find as many relevant images as possible

- Fifty user needs (topics):
  - Expressed with a short (title) and longer (narrative) textual description
  - Also expressed with an example relevant image (QBE)
  - Titles translated into 5 European languages (by Sheffield) and Chinese (by NTU)
- Two retrieval challenges
  - Matching textual queries to visual documents (*use captions*)
  - Matching non-English queries to English captions (*use translation*)
- Essentially a bilingual CLIR task
- No retrieval constraints specified

Creating the test collection



Evaluation

- Evaluation based on most stringent relevance set (strict intersection)
- Compared systems using
  - MAP across all topics
  - Number of topics with no relevant image in the top 100
- 4 participants evaluated (used captions only):
  - NTU – Chinese->English, manual and automatic, Okapi and dictionary-based translation, focus on proper name translation
  - Daedalus – all->English (except Dutch and Chinese), Xapian and dictionary-based + on-line translation, Wordnet query expansion, focus on indexing query and ways of combining query terms
  - Surrey – all->English (except Chinese), SoCIS system and on-line translation, Wordnet expansion, focus on query expansion and analysis of topics
  - Sheffield – all->English, GLASS (BM25) and Systran translation, no language-specific processing, focus on translation quality

Results

- Surrey had problems
- NTU obtained highest Chinese results
  - approx. 51% mono and 12 failed topics (NTUiaCoP)
- Sheffield obtained highest
  - Italian: 72% mono and 7 failed topics
  - German: 75% mono and 8 failed topics
  - Dutch: 69% mono and 7 failed topics
  - French: 78% mono and 3 failed topics
- Daedalus obtained highest
  - Spanish: 76% mono and 5 failed topics (QTdoc)
  - Monolingual: 0.5718 and 1 failed topic (Qor)
- For more information ... see the ImageCLEF working notes