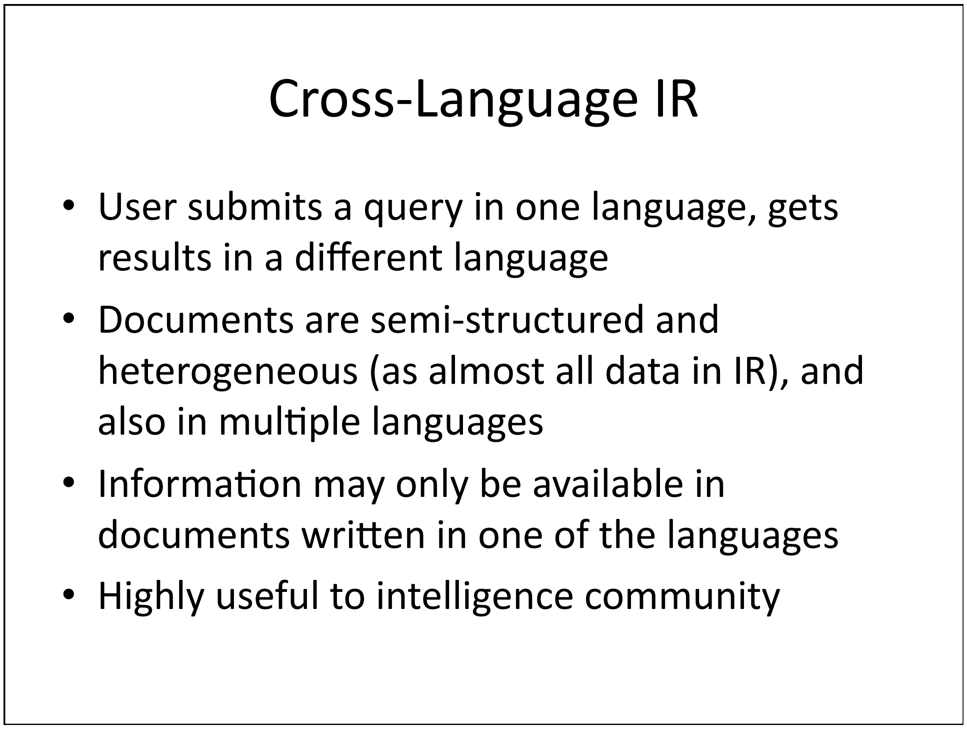
# Cross-Language IR

CISC489/689-010, Lecture #23

Monday, May 11th

Ben Carterette



## Cross-Language IR

- User submits a query in one language, gets results in a different language
- Documents are semi-structured and heterogeneous (as almost all data in IR), and also in multiple languages
- Information may only be available in documents written in one of the languages
- Highly useful to intelligence community

# Approaches to CLIR

- Translate the documents into the users' language, and let the users submit queries in their own language
- Translate the users' queries into target language(s) and use the translated query for retrieval
- Translate both queries and documents to an "intermediate" language

# Automatic Translation

- What are some approaches to automatic translation?
  - Language-to-language dictionaries
- Languages do not translate precisely
  - One word with several meanings in one language might translate to several different words in the other
  - Many words with the same meaning might all translate to a single word
  - A word in one language might only be expressible as a phrase in another (or vice-versa)
  - etc…

# Example

- English queries to retrieve Spanish documents
- System works by translating query to Spanish
- Query: “bank fraud”
- Translations of “bank”:
  - *Orilla* (river bank)
  - *Terraplen* (bank of earth)
  - *Banco* (bank of clouds)
  - *Bateria* (bank of lights)
  - *Banco* (financial institution)
  - *Banca* (casino bank)
- Translations of “fraud”:
  - *Impostor* (fraudulent person)
  - *Fraude* (deception)
- How would a dictionary-based system know which pair of translations to use?
- Possibly correct translation:
  - *Fraude bancario*

# Statistical Approach

- Instead of trying to translate directly, apply statistical methods
- Learn “translation probabilities” P(f | e) – probability of translating string e in language E to string f in language F
- E.g.:
  - P(orilla fraude | bank fraud), P(orilla impostor | bank fraud), P(banco fraude | bank fraud), …

## Cross-Language Language Model

- Recall query-likelihood language model:

$$P(Q|D) = \prod_{q \in Q} P(q|D) = \prod_{q \in Q} (1 - \alpha_D) \frac{tf_{qD}}{|D|} + \alpha_D \frac{ctf_q}{|C|}$$

- Let's adapt this to cross-language retrieval using statistical translation

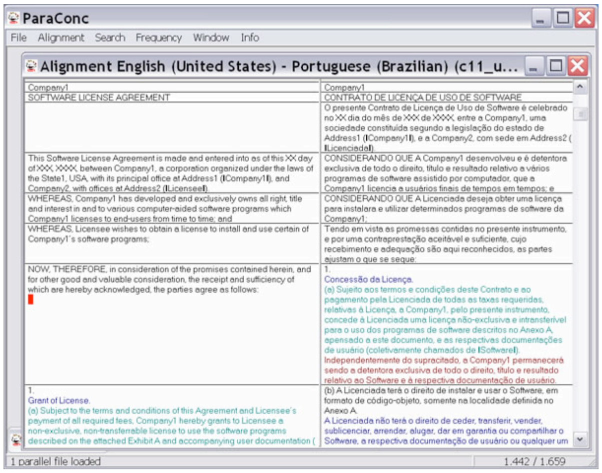
$$P(Q_f|D_e) \quad = \quad \prod_{q_f \in Q_f} P(q_f|D_e)$$

## Translation Model

- What is $P(q_f \mid t_e)$?
- The *translation model*: probability of translating word $t_e$ in language E to word $q_f$ in language F
- Where does it come from?
  - Maybe a dictionary approach: every possible translation of $t_e$ has equal probability
  - e.g. P(orilla | bank) = P(banco | bank) = P(banca | bank) = …

# Statistical Translation Model

- An alternative approach: *parallel corpora*



| | |
|---|---|
| Company1 | Company1 |
| SOFTWARE LICENSE AGREEMENT | CONTRATO DE LICENÇA DE USO DE SOFTWARE |
| | O presente Contrato de Licença de Uso de Software é celebrado no XX dia do mês de XXX de XXXX, entre a Company1, uma sociedade constituída segundo a legislação do estado de Address1 (ICompany1I), e a Company2, com sede em Address2 (ILicenciadaI). |
| This Software License Agreement is made and entered into as of this XX day of XXX, XXXX, between Company1, a corporation organized under the laws of the State1, USA, with its principal office at Address1 (ICompany1I), and Company2, with offices at Address2 (ILicenseeI). | CONSIDERANDO QUE A Company1 desenvolveu e é detentora exclusiva de todo o direito, título e resultado relativo a vários programas de software assistido por computador, que a Company1 licencia a usuários finais de tempos em tempos; e |
| WHEREAS, Company1 has developed and exclusively owns all right, title and interest in and to various computer-aided software programs which Company1 licenses to end-users from time to time; and | CONSIDERANDO QUE A Licenciada deseja obter uma licença para instalara e utilizar determinados programas de software da Company1; |
| WHEREAS, Licensee wishes to obtain a license to install and use certain of Company1's software programs; | Tendo em vista as promessas contidas no presente instrumento, e por uma contraprestação aceitável e suficiente, cujo recebimento e adequação são aqui reconhecidos, as partes ajustam o que se segue: |
| NOW, THEREFORE, in consideration of the promises contained herein, and for other good and valuable consideration, the receipt and sufficiency of which are hereby acknowledged, the parties agree as follows: | 1.<br>Concessão da Licença.<br>(a) Sujeito aos termos e condições deste Contrato e ao pagamento pela Licenciada de todas as taxas requeridas, relativas à Licença, a Company1, pelo presente instrumento, concede à Licenciada uma licença não-exclusiva e intransferível para o uso dos programas de software descritos no Anexo A, apensado a este documento, e as respectivas documentações de usuário (coletivamente chamados de ISoftwareI).<br>Independentemente do supracitado, a Company1 permanecerá sendo a detentora exclusiva de todo o direito, título e resultado relativo ao Software e à respectiva documentação de usuário. |
| 1.<br>Grant of License.<br>(a) Subject to the terms and conditions of this Agreement and Licensee's payment of all required fees, Company1 hereby grants to Licensee a non-exclusive, non-transferrable license to use the software programs described on the attached Exhibit A and accompanying user documentation ( | (b) A Licenciada terá o direito de instalar e usar o Software, em formato de código-objeto, somente na localidade definida no Anexo A.<br>A Licenciada não terá o direito de ceder, transferir, vender, sublicenciar, arrendar, alugar, dar em garantia ou compartilhar o Software, a respectiva documentação de usuário ou qualquer um |

# Statistical Translation with Parallel Corpora

- Parallel corpora consist of documents in two or more languages that are known to be translations of one another
- The parallel copora are *aligned*: string e and string f are marked as translations of each other
- We can use these alignments to estimate a translation model

# Translation Model

- To estimate P($q_f$ | $t_e$), count the number of aligned string pairs (e, f) such that $t_e$ is a word in e and $q_f$ is a word in f
- Divide by the total number of strings in language e that contain $t_e$

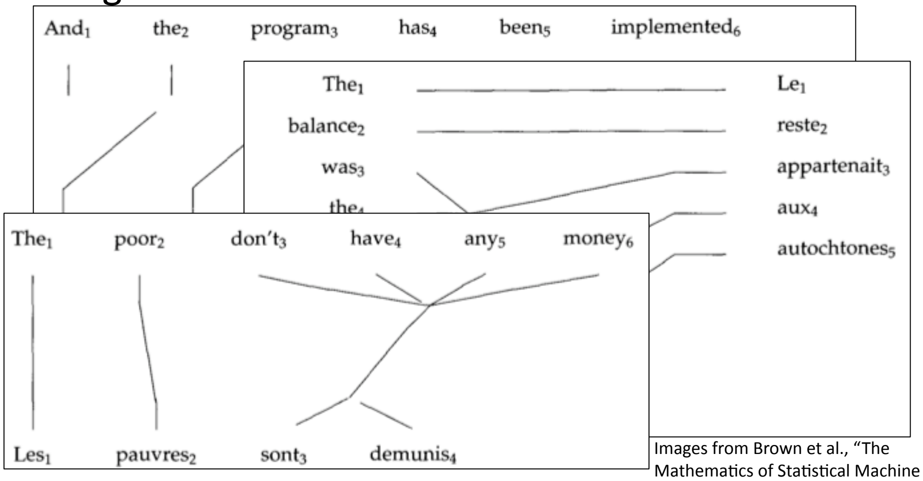
$$P(q_f|t_e) = \frac{|\{(e,f)|t_e \in e \text{ and } q_f \in f\}|}{|\{e|t_e \in e\}|}$$

# Simple Alignment Example

- English sentence: “The objective was clear: arrest and extradite to Mexico the woman against whom they had charged for fraud to a recognized banking institution.”
- Spanish sentence: “El objetivo era claro: detener a la mujer y enviarla de regreso a México pues habían cargos en su contra por fraude a una reconocida institución bancaria.”
- Every pair of words in these two sentences will have some translation probability
- Over many sentences, the highest probabilities will be the pairs of words that are most closely related

# Alignments

- Alignments can be much more detailed


And$_1$ the$_2$ program$_3$ has$_4$ been$_5$ implemented$_6$
The$_1$
balance$_2$
was$_3$
the$_4$
Le$_1$
reste$_2$
appartenait$_3$
aux$_4$
autochtones$_5$
The$_1$ poor$_2$ don't$_3$ have$_4$ any$_5$ money$_6$
Les$_1$ pauvres$_2$ sont$_3$ demunis$_4$

Images from Brown et al., "The Mathematics of Statistical Machine Translation"

# Parallel Corpora

- Where do we get parallel corpora?
  - Find documents that we know to be translations
  - Canadian Hansard: transcripts of Canadian parliamentary debates in both English and French
  - European Union law in 22 languages
- Anything that's not law-related?
  - Wikipedia articles in different languages.. Not necessarily translations though

# CLIR Experiments

- CLIR track ran at TREC from 1998 through 2002
- Languages used include English, German, French, Italian, Chinese, and Arabic
- Other issues in CLIR:
  - Segmentation, stemming, stopping, phrases require different approaches in different languages
  - I am going to focus on high-level problem

# CLIR Experiments

- In 2001 and 2002, the main CLIR task was English queries to retrieve Arabic documents
- Documents: 383,872 news articles from Agence France Press from 1994-2000
- Information needs: 25 queries, descriptions, and narratives in English by native Arabic speakers
  - Translated into Arabic and French as well
- Participating sites could do CLIR (English to Arabic or French to Arabic) or normal IR (Arabic to Arabic)

# Example Topic

<num> Number: AR26
<title> مجلس المقاومة الوطني الكردستاني

<desc> Description:

كيف ينظر مجلس المقاومة الوطنية الى الإستقلال المحتمل للاكراد؟

<narr> Narrative:

الموضوع يتضمن نصوص متعلقة بتحركات مجلس المقاومة الوطنية ، مقالات تتحدث عن قيادة اوجلان ضمن جهود الاكراد للاستقلال .

<num> Number: AR26
<title> Kurdistan Independence

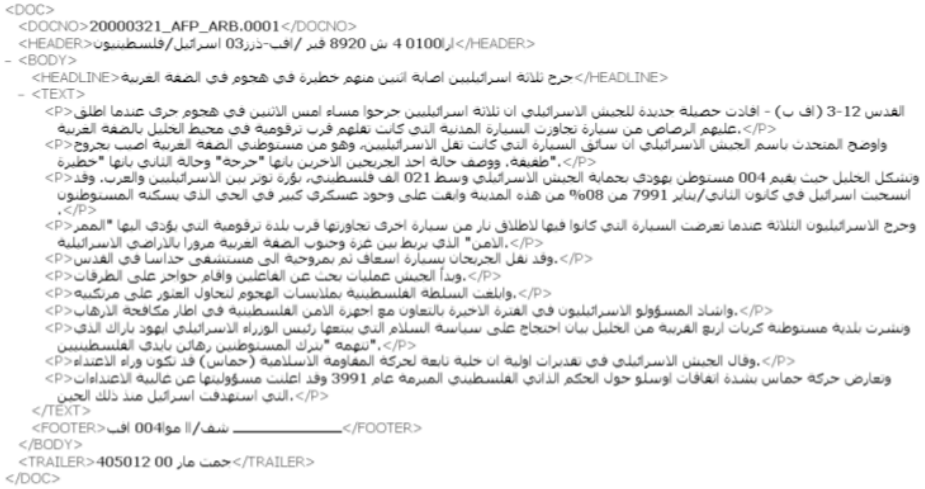
<desc> Description:

How does the National Council of Resistance relate to the potential independence of Kurdistan?

<narr> Narrative:

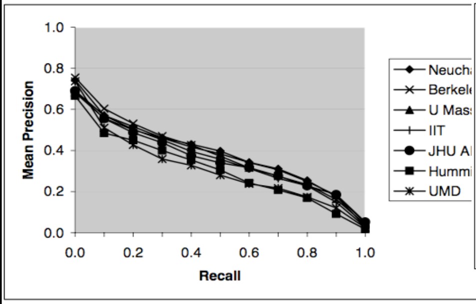
Articles reporting activities of the National Council of Resistance are considered on topic. Articles discussing Ocalan's leadership within the context of the Kurdish efforts toward independence are also considered on topic.

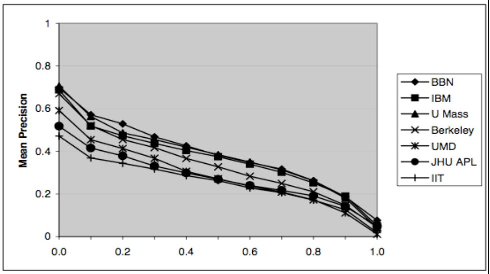
# Example Document



```
<DOC>
  <DOCNO>20000321_AFP_ARB.0001</DOCNO>
  <HEADER>ار0100 4 ش 8920 فبر /اف-ذزز03 اسرائيل/فلسطينيون</HEADER>
- <BODY>
    <HEADLINE>جرح ثلاثة اسرائيليين اصابة اثنين منهم خطيرة في هجوم في الضفة الغربية</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 ترقومية 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 ترقومية التي يؤدي اليها "الممر
       الامن" الذي يربط بين غزة وجنوب الضفة الغربية مرورا بالاراضي الاسرائيلية.</P>
      <P>وقد نقل الجريحان بسيارة اسعاف ثم 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 مستوطنة كريات اربع القريبة من الخليل بيان احتجاج على سياسة السلام التي يتبعها رئيس الوزراء الاسرائيلي ايهود باراك الذي
       تتهمه "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>اف 004مو/اا شف ________________</FOOTER>
  </BODY>
  <TRAILER>جمت مار 00 405012</TRAILER>
</DOC>
```

# Results





Cross-lingual (English/French to Arabic)

Monolingual (Arabic to Arabic)

- BBN, Umass, IBM used statistical models
- Umass performance on cross-language is roughly equal to performance on monolingual!

Plots from Oard & Gey, "The TREC-2002 Arabic/English CLIR Track"

# Analysis

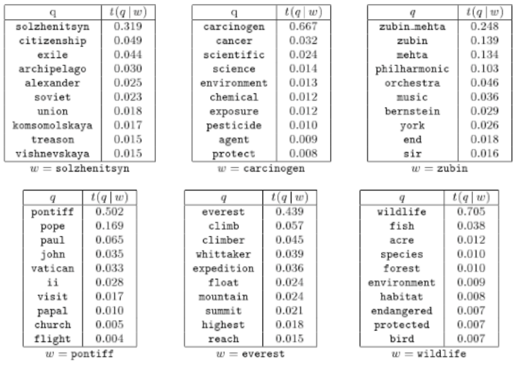
- The translation model is imperfect
  - It assigns probabilities to almost every pair of words
  - There are many errors in translation
- So how could cross-lingual be almost as good as monolingual?
- Hypotheses:
  - Translation process disambiguates some terms
  - Translation process smooths query models

# IR as Statistical Translation

- What if we view IR as a translation process?
  - User inputs query in English, system does “cross-language” retrieval from user-English to system-English
  - This may account for users not using the right keywords in their queries
- There is no natural translation model, so one must be simulated
- Berger & Lafferty, SIGIR 1999

# IR Translation Model

- Generate a translation model by aligning simulated queries to relevant documents



| $q$ | $t(q \mid w)$ |
|---|---|
| solzhenitsyn | 0.319 |
| citizenship | 0.049 |
| exile | 0.044 |
| archipelago | 0.030 |
| alexander | 0.025 |
| soviet | 0.023 |
| union | 0.018 |
| komsomolskaya | 0.017 |
| treason | 0.015 |
| vishnevskaya | 0.015 |

$w =$ solzhenitsyn

| $q$ | $t(q \mid w)$ |
|---|---|
| carcinogen | 0.667 |
| cancer | 0.032 |
| scientific | 0.024 |
| science | 0.014 |
| environment | 0.013 |
| chemical | 0.012 |
| exposure | 0.012 |
| pesticide | 0.010 |
| agent | 0.009 |
| protect | 0.008 |

$w =$ carcinogen

| $q$ | $t(q \mid w)$ |
|---|---|
| zubin_mehta | 0.248 |
| zubin | 0.139 |
| mehta | 0.134 |
| philharmonic | 0.103 |
| orchestra | 0.046 |
| music | 0.036 |
| bernstein | 0.029 |
| york | 0.026 |
| end | 0.018 |
| sir | 0.016 |

$w =$ zubin

| $q$ | $t(q \mid w)$ |
|---|---|
| pontiff | 0.502 |
| pope | 0.169 |
| paul | 0.065 |
| john | 0.035 |
| vatican | 0.033 |
| ii | 0.028 |
| visit | 0.017 |
| papal | 0.010 |
| church | 0.005 |
| flight | 0.004 |

$w =$ pontiff

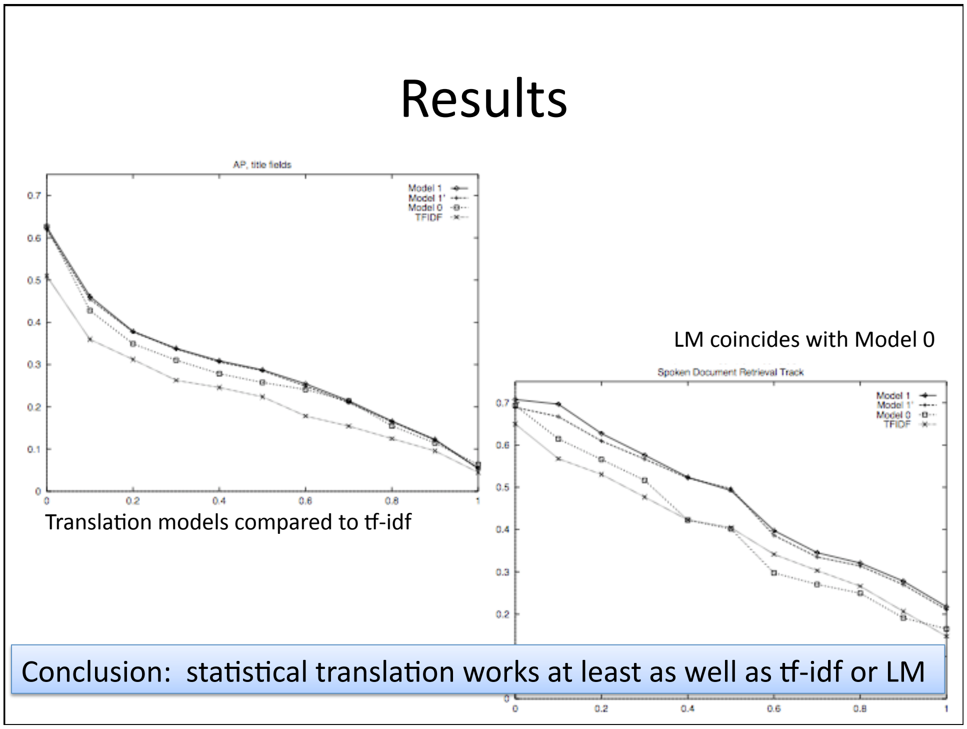
| $q$ | $t(q \mid w)$ |
|---|---|
| everest | 0.439 |
| climb | 0.057 |
| climber | 0.045 |
| whittaker | 0.039 |
| expedition | 0.036 |
| float | 0.024 |
| mountain | 0.024 |
| summit | 0.021 |
| highest | 0.018 |
| reach | 0.015 |

$w =$ everest

| $q$ | $t(q \mid w)$ |
|---|---|
| wildlife | 0.705 |
| fish | 0.038 |
| acre | 0.012 |
| species | 0.010 |
| forest | 0.010 |
| environment | 0.009 |
| habitat | 0.008 |
| endangered | 0.007 |
| protected | 0.007 |
| bird | 0.007 |

$w =$ wildlife

**Figure 2.** *Sample translation probabilities after EM training on synthetic data.*

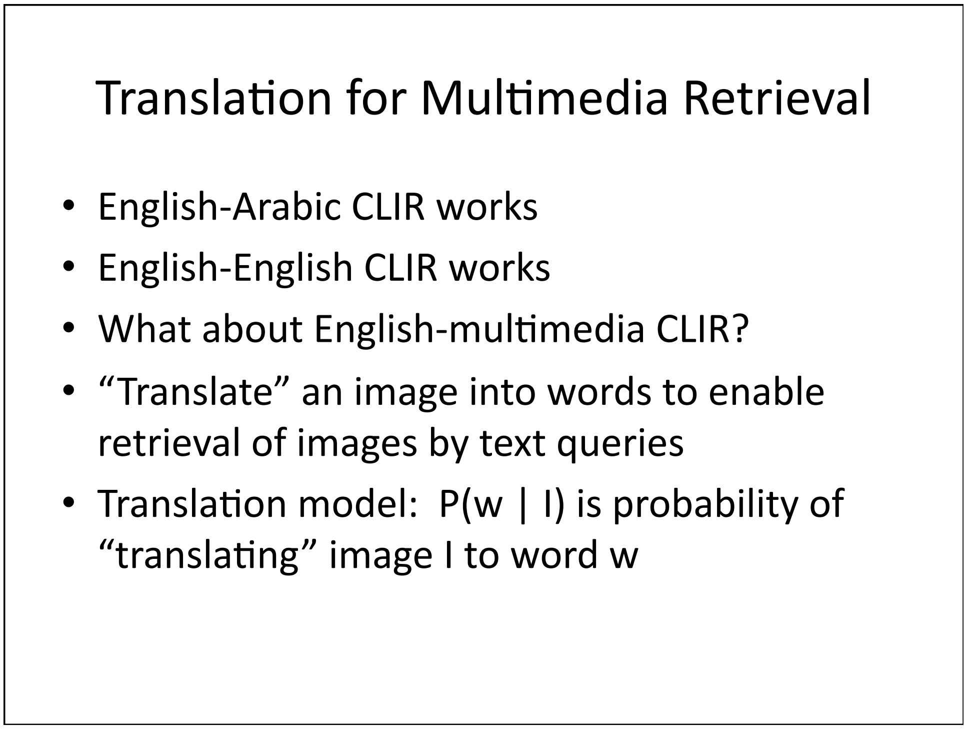


## Results

Translation models compared to tf-idf

LM coincides with Model 0

Conclusion: statistical translation works at least as well as tf-idf or LM



## Translation for Multimedia Retrieval

- English-Arabic CLIR works
- English-English CLIR works
- What about English-multimedia CLIR?
- “Translate” an image into words to enable retrieval of images by text queries
- Translation model: P(w | I) is probability of “translating” image I to word w

# Image Translation Model

- Estimate P(w | I) requires two things:
  - A feature-based representation of the image
  - A set of words that "align" with the image
- Use image segmentation and clustering to form a representation of images
- Use image captions to align words to image

# Image Representation: "Blobs"

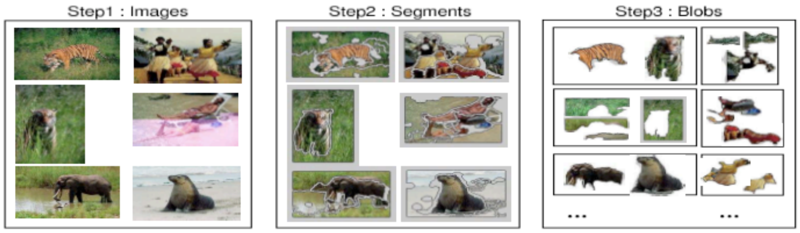


Figure 2: Image preprocessing: Step 2 shows the segmentation results from a typical segmentation algorithm (Blobworld) The clusters in step 3 are manually constructed to show the concept of blobs. Both the segmentation and the clustering often produce semantically inconsistent segments (breaking up the tiger) and blobs (seals and elephants in the same blob) .

From Jeon et al., "Automatic Image Annotation and Retrieval Using Cross-Media Relevance Models"

# Cross-Media Relevance Model

- Retrieval is by query-likelihood P(Q | I)

$$
\begin{aligned}
P(Q|I) &= \prod_{q \in Q} P(q|I) \\
&\approx \prod_{q \in Q} P(q|b_1, ..., b_m) \\
&\propto \prod_{q \in Q} \sum_{J \in C} P(q|J)P(J) \prod_{i=1}^{m} P(b_i|J)
\end{aligned}
$$

C is the collection of images, J is an image in C, and $b_1 \ldots b_m$ are "blobs"

# Example Results

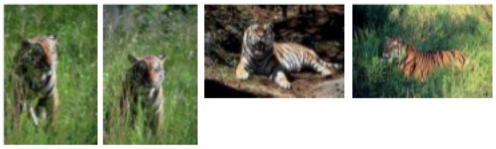


**Figure 7: Retrieval (DRCMRM) in response to the text query "tiger".**

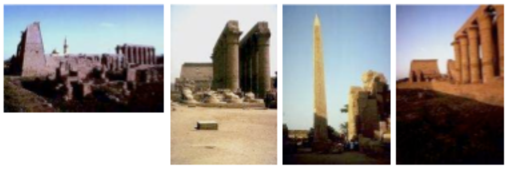


**Figure 8: Retrieval (DRCMRM) in response to the text query "pillar". Note the pillar(s) in each image**

From Jeon et al., "Automatic Image Annotation and Retrieval Using Cross-Media Relevance Models"

# Machine Translation

- Machine translation (MT) is a problem in NLP/ computational linguistics
- The goal is to automatically translate text in one language to another
- Different from CLIR with query translation model in that the CLIR model does not require a "coherent" translation of the query
  - CLIR essentially uses every possible translation
- Machine translation should provide a single "good" translation that is human-readable

# Statistical MT

- Though MT and CLIR are different problems, the statistical approaches are very similar
- IBM developed several statistical models for MT
  - "A statistical approach to machine translation", Brown et al. 1990
  - CLIR models based on IBM's models

# IBM Models

- Basic idea: to translate a sentence f in language F to a sentence e in language E, estimate P(e | f) using Bayes Rule

$$P(e|f) = \frac{P(f|e)P(e)}{P(f)}$$

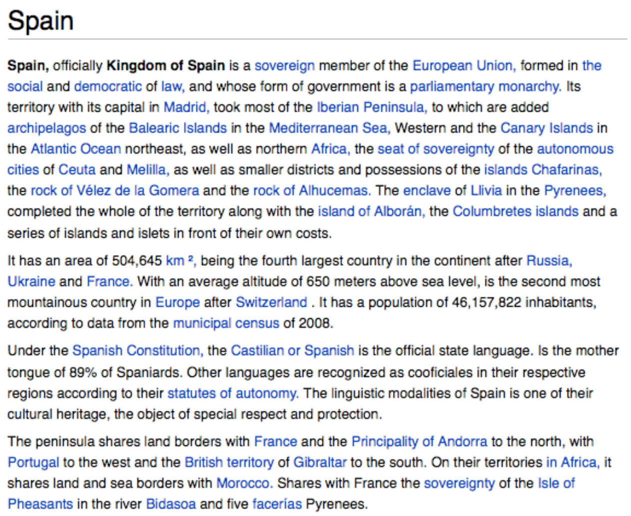
- The "right" translation is the one with highest probability

$$\widehat{e} = \arg\max_{e} P(f|e)P(e)$$

# IBM Models

- The key is estimating P(f | e)
- Brown et al. presented five different models
  - Increasingly complicated, require a lot of training data in the form of parallel aligned corpora
- Google machine translation is based on alignment and IBM models, but also based on very large amounts of unaligned data

# Google Machine Translation



## Spain

**Spain,** officially **Kingdom of Spain** is a sovereign member of the European Union, formed in the social and democratic of law, and whose form of government is a parliamentary monarchy. Its territory with its capital in Madrid, took most of the Iberian Peninsula, to which are added archipelagos of the Balearic Islands in the Mediterranean Sea, Western and the Canary Islands in the Atlantic Ocean northeast, as well as northern Africa, the seat of sovereignty of the autonomous cities of Ceuta and Melilla, as well as smaller districts and possessions of the islands Chafarinas, the rock of Vélez de la Gomera and the rock of Alhucemas. The enclave of Llivia in the Pyrenees, completed the whole of the territory along with the island of Alborán, the Columbretes islands and a series of islands and islets in front of their own costs.

It has an area of 504,645 km $^2$, being the fourth largest country in the continent after Russia, Ukraine and France. With an average altitude of 650 meters above sea level, is the second most mountainous country in Europe after Switzerland . It has a population of 46,157,822 inhabitants, according to data from the municipal census of 2008.

Under the Spanish Constitution, the Castilian or Spanish is the official state language. Is the mother tongue of 89% of Spaniards. Other languages are recognized as cooficiales in their respective regions according to their statutes of autonomy. The linguistic modalities of Spain is one of their cultural heritage, the object of special respect and protection.

The peninsula shares land borders with France and the Principality of Andorra to the north, with Portugal to the west and the British territory of Gibraltar to the south. On their territories in Africa, it shares land and sea borders with Morocco. Shares with France the sovereignty of the Isle of Pheasants in the river Bidasoa and five facerías Pyrenees.

Google’s translation of the Spanish Wikipedia page for Spain (http://es.wikipedia.org/wiki/Espana)