Best Practices in Multilingual Search Assistance

JULIO GONZALO
NLP.UNED.ES, MADRID
MLIA TECHNOLOGY DAY
BERLIN, 8 DECEMBER 2009
An example

English translation

China Air Force F-loaded column 14: F22 of the shell, F-11 core, car soldiers _ Sina blog
7 soldiers blog, car, seven soldiers with vehicles, recently installed on China's F-14 column news a lot lately, I believe that these messages are not something out of nothing. I believe that China has the capability to produce such a F-14: F22 of the shell, F-11 core.
blog.sina.com.cn/s/blog_5ce33beb0100emaa.html - 23k - Cached

Original Simplified Chinese - Hide Simplified Chinese results

中国空军列装歼14：F22的外壳，歼11的内核 _ 七连汽车兵_新浪博客
七连汽车兵的博客,七连汽车兵 最近，关于中国列装歼14的消息满天飞。笔者认为这些消息并非无中生有。笔者认为中国有能力制造这样一种歼14：F22的外壳，歼11的内核。
blog.sina.com.cn/s/blog_5ce33beb0100emaa.html - 23k - 网页快照

Posted eye-catching HP DV2-1005AX shell paste (smartec HP DV2-1005AX shell paste) notebook
Eye-catching cover affixed HP DV2-1005AX Notebook paste beauty products, including eye-catching cover affixed HP DV2-1005AX Notebook beauty paste quotes, pictures, product parameters, Function, User reviews, product forums, concerned about the degree, eye-catching paste...
product.pconline.com.cn/nbmakeup/smartec/347936.html - 46k - Cached

醒目贴惠普DV2-1005AX外壳贴(smartec 惠普DV2-1005AX外壳贴)笔记本美容 ...
醒目贴惠普DV2-1005AX外壳贴笔记本美容产品信息，包括醒目贴惠普 DV2-1005AX外壳贴 笔记本美容报价、图片、产品参数、功能介绍、网友评论、产品论坛、关注度，醒目贴 ...
product.pconline.com.cn/nbmakeup/smartec/347936.html - 46k - 网页快照

Mobile Shell - mobile phone housings - mobile phone housings do Alibaba - Shenzhen Tongda Golden Metal

手机外壳-手机外壳-手机外壳尽在阿里巴巴-深圳市通达金五金制品有限公司
Sources of evidence

iCLEF (interactive CLEF track)
• Outcome of scientific experimentation

TrebleCLEF Workshop on *Best Practices for the development of Multilingual Information Access systems: The User Perspective* (Segovia, June 2008)
• Advice from relevant user communities
Cross-Language Information Access Assistance

Aspects

Document selection
Query formulation & translation
Feedback and query refinement

Tasks

Text Retrieval
Image Retrieval
Question Answering
# Cross-Language Information Access Assistance

<table>
<thead>
<tr>
<th>Aspects</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document selection</td>
<td>Text Retrieval</td>
</tr>
<tr>
<td>Query formulation &amp; translation</td>
<td>Image Retrieval</td>
</tr>
<tr>
<td>Feedback and query refinement</td>
<td>Question Answering</td>
</tr>
</tbody>
</table>
Document selection: Is this document relevant?

Query bridge and tunnel construction for the Beijing-Kowloon Railroad
Title coal eastward
Date 99/07/07

Source: Oard 2000
Cross-Language Document Selection

High quality translations pay off.

Cross-Language summaries preferable to full MT.

Translating the document collection (or even document summaries) at index time pays off.

Offer flexibility about how to organize results (by language or merged)

If translation is not available, show metadata and/or perform approximate translation

Alert the user when there is relevant information in other languages, and about the quality of MT
<table>
<thead>
<tr>
<th>Aspects</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document selection</td>
<td>Text Retrieval</td>
</tr>
<tr>
<td>Query formulation &amp; translation</td>
<td>Image Retrieval</td>
</tr>
<tr>
<td>Feedback and query refinement</td>
<td>Question Answering</td>
</tr>
</tbody>
</table>
Do not show assisted query translation by default
Include option to fix translation (I)
Include option to fix translation (II)
Include option to fix translation (II)
But help the user choosing the right translation!
Even preferable: monolingual assistance

(A) Clicking on best-aligned phrases incorporates them to the query.
Indirect assistance works well!
Design batch & interactive CLIR together
Design batch & interactive CLIR together
Query Translation and Refinement wrap-up

Include user-assisted query translation facilities...

... but do not show them by default

Indirect user-assisted translation (without target-language examination) is preferable

Link structured sources that help mapping the meaning of the query (Wikipedia, named entities, KWIC, synonyms...)

Query translation, document translation and assisted query translation must be designed to fit together
Personalization: language profiles

Flickling - An online game for searching Flickr across language boundaries

Mi perfil

Nombre de usuario: julio.gonzalo@gmail.com
Equipo: testing

Lengua nativa: Español
Lengua de la interfaz: Español
Lenguas desconocidas: Alemán, Holandés
Lenguas pasivas: Francés, Italiano
Lenguas activas: Inglés

Ranking: 191
Puntuación: 20
Precision: 0.5
Tiempo medio: 65.62
Wrap-up

- Interactive aspects of Multilingual Information Access are key to successful system building
- Check TrebleCLEF best practice recommendations (pick up one copy of the deliverable now!)
- Be the first to build a multilingual search assistant that complies with TrebleCLEF recommendations!
Backup slides
Case study 1: Google Translated Search

**Translated Search**

Search for: shell  Translated to: 外壳 - Not quite right? Edit
My language: English  Search pages written in: Chinese (Simplified) -
Translate and Search

**Translated results from Simplified Chinese web pages**

Results 1 - 10 of about 10,200,000 for 外壳.

**English translation**

*China Air Force F-loaded column 14: F22 of the shell, F-11 core, car soldiers __ 7 Sina blog*
7 soldiers blog, car, seven soldiers with vehicles, recently installed on China's F-14 column news a lot lately. I believe that these messages are not something out of nothing. I believe that China has the capability to produce such a F-14: F22 of the shell, F-11 core.
blog.sina.com.cn/s/blog_5ce33be0100ema.html - 23k - Cached

**Original Simplified Chinese**

*中国空军列装歼14·F22的外壳·歼11的内核·七连汽车兵__新浪博客*
七连胜汽车兵的博客;七连胜汽车兵, 最近，关于中国列装歼14的消息满天飞。笔者认为这些消息并非无中生有。笔者认为中国有能力制造这样一种歼14：F22的外壳，歼11的内核。
blog.sina.com.cn/s/blog_5ce33be0100ema.html - 23k - 网页快照

**Posted eye-catching HP DV2-1005AX shell paste (smartec HP DV2-1005AX shell paste) notebook Beauty...**
Eye-catching cover affixed HP DV2-1005AX Notebook paste beauty products, including eye-catching cover affixed HP DV2-1005AX Notebook paste beauty quotes, pictures, product parameters. Function, User reviews, product forums, concerned about the degree, eye-catching paste...
product.pconline.com.cn/nbmakeup/smartec/347936.html - 46k - Cached

**醒目貼惠普DV2-1005AX外壳貼(smartec 惠普DV2-1005AX外壳贴)笔记本美容...**
醒目貼惠普DV2-1005AX外壳貼笔记本美容产品信息，包括醒目貼惠普DV2-1005AX外壳貼 笔记本美容报价、图片、产品参数、功能介绍、网友评论、产品论坛、关注度，醒目貼...
product.pconline.com.cn/nbmakeup/smartec/347936.html - 46k - 网页快照

**Mobile Shell - mobile phone housings - mobile phone housinas do Alibaba - Shenzhen Tonada Golden Metal**

**手機外壳-手机外壳-手机外壳尽在阿里巴巴-深圳市通达金五金制品有限公司**
Translated Search

Search for: shell
My language: English
Translated to: 外壳
Search pages written in: Chinese (Simplified)

Translate and Search

Translated results from Simplified Chinese web pages

English translation

China Air Force F-loaded column 14: F22 of the shell, F-11 core, car soldiers __ 7 Sina blog
7 soldiers blog, car, seven soldiers with vehicles, recently installed on

Original Simplified Chinese - Hide:

中国空军列装歼14：F22的外壳，歼11核心，汽车士兵 __ 7 新浪博客
七连汽车兵的博客,七连汽车兵,最近，关于中
<table>
<thead>
<tr>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-quality Machine Translation</td>
<td>No consideration of user’s language skills</td>
</tr>
<tr>
<td>Cross-Language Summaries (snippets) for document selection</td>
<td>Only one target language at once</td>
</tr>
<tr>
<td>Interactive query translation hidden by default</td>
<td>No assistance for query translation</td>
</tr>
</tbody>
</table>
Case study 2: Flickling

Flickling - An online game for searching Flickr across language boundaries - Mozilla Firefox

http://soporte.1lsi.unic.es/flickling/index.html?locale=es#%7B%22start_index%22%3A0%2C%22num_results%22%3A20%2C%22%7D

julio.gonzalo@gmail.com | Mi perfil | Mostrar puntuaciones | Ayuda

Flickling - An online game for searching Flickr across language boundaries

Encuentra esta imagen | monolingüe | multilingüe

violin strauss estatua

escribo en traduce mi consulta a

Español | DE | EN | ES | FR | IT | NL

Resultados 1-16 de 16 para violin strauss estatua (Mostrar consulta a Flickr)

Tal vez quieras intentar con: music, concert, instrument, strings, musician, fiddle, guitar, live, bw, street... [mostrar todo]

Strauss

vienna, park, parque, music, classic, austria, golden, violin, viena, musica... [mostrar todo]

P4302599
Flickling: Relevance feedback

Tal vez quieras intentar con: music, concert, instrument, strings, musician, fiddle, guitar, live, bw, street...

Strauss

vienna, park, parque, music, classic, austria, golden, violin, viena

P4302599

vienna, wien, music, classic, austria, golden, europe, violina

P4302607

**Flickling**

An online game for searching Flickr across language boundaries

<table>
<thead>
<tr>
<th>Encuentra esta imagen</th>
<th>monolingüe</th>
<th>multilingüe</th>
</tr>
</thead>
<tbody>
<tr>
<td>violin strauss estatua</td>
<td></td>
<td></td>
</tr>
<tr>
<td>escribo en traduce mi consulta a</td>
<td>DE EN ES FR IT NL</td>
<td></td>
</tr>
</tbody>
</table>

**Traducciones**

<table>
<thead>
<tr>
<th>Español</th>
<th>violin</th>
<th>strauss estatua</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alemán</td>
<td>violine, geige, violinist</td>
<td>[añadir] standbild</td>
</tr>
<tr>
<td>Inglés</td>
<td>violin</td>
<td>[añadir] [añadir]</td>
</tr>
<tr>
<td>Francés</td>
<td>violon, violoniste</td>
<td>[añadir] statue</td>
</tr>
<tr>
<td>Italiano</td>
<td>violino, violinista</td>
<td>[añadir] statua</td>
</tr>
<tr>
<td>Holandés</td>
<td>violist</td>
<td>[añadir] beeld</td>
</tr>
</tbody>
</table>

**Resultados 1-16 de 16 para violin strauss estatua**

(Mostrar consulta a Flickr)

Tal vez quieras intentar con: music, concert, instrument, strings, musician, fiddle, guitar, live, bw, street...

[mostrar todo]

**Strauss**
Case study 2: Flickling

<table>
<thead>
<tr>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Multiple language search</td>
<td>• No assistance to choose between alternative translations</td>
</tr>
<tr>
<td>• Assisted query translation hidden by default</td>
<td>• No option to see results separated by language</td>
</tr>
<tr>
<td>• Language skills in user’s profile</td>
<td></td>
</tr>
</tbody>
</table>
Case study: MultiMatch
### Case study 3: MultiMatch

<table>
<thead>
<tr>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Multiple language search</td>
<td>• Mismatch between automatic and assisted query translation machinery</td>
</tr>
<tr>
<td>• Integrated with monolingual search</td>
<td>• No assistance to choose between alternative translations</td>
</tr>
<tr>
<td>• Assisted query translation hidden by default</td>
<td>• No adaptation to user’s language skills</td>
</tr>
<tr>
<td>• Option to see results merged or by language</td>
<td>•</td>
</tr>
</tbody>
</table>
## Query Translation and Refinement

<table>
<thead>
<tr>
<th>Query translation &amp; refinement</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Include user-assisted query translation facilities.</td>
<td>(He et al. 2003), (Dorr et al. 2004)</td>
</tr>
<tr>
<td>Hide user-assisted query translation by default, make it available when things go wrong.</td>
<td>(Petrelli et al. 2003) and most observational studies at iCLEF; Treble CLEF User Communities Workshop (Gonzalo et al. 2009).</td>
</tr>
<tr>
<td>Indirect user-assisted query translation that does not involve inspecting foreign-language terms is preferable.</td>
<td>(López-Ostenero et al. 2005a), feedback via suggestion of noun phrases related to the query and taken from the document collection.</td>
</tr>
<tr>
<td>Document translation and query translation/formulation/refinement facilities must be consistently designed to fit together.</td>
<td>(López-Ostenero et al. 2005a)</td>
</tr>
<tr>
<td>If possible, link structured sources that help mapping the meaning of the query, e.g. named entities, Wikipedia entries.</td>
<td>Treble CLEF User Communities Workshop (Gonzalo et al. 2009)</td>
</tr>
</tbody>
</table>
# Cross-Language Document Selection

<table>
<thead>
<tr>
<th>Recommendations (for user interface)</th>
<th>Based On</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cross-Language Document Selection</strong></td>
<td></td>
</tr>
<tr>
<td>Try to offer high-quality translations to the user.</td>
<td>(Oard et al. 2004)</td>
</tr>
<tr>
<td>Cross-language document summaries are often preferable to full machine translation of docs.</td>
<td>(Oard et al. 2004), (Llopis et al. 2003), (Navarro et al. 2004), (Richardson 2007)</td>
</tr>
<tr>
<td>If possible, translating the whole document collection at index time pays off. Translating a document summary can also work.</td>
<td>(Ostenero et al. 2004), (Oard 2009), Treble CLEF User Communities Workshop (Gonzalo et al. 2009)</td>
</tr>
<tr>
<td>Interfaces should be flexible about how to organize results (by target language or merged).</td>
<td>Treble CLEF User Communities Workshop (Gonzalo et al. 2009)</td>
</tr>
<tr>
<td>If translation is not available, try to show metadata, translate into the user’s second language, or perform approximate translation.</td>
<td>Treble CLEF User Communities Workshop (Gonzalo et al. 2009)</td>
</tr>
<tr>
<td>The system should alert the user (i) whenever there is more information available in other target languages, (ii) about the quality of MT.</td>
<td>Treble CLEF User Communities Workshop (Gonzalo et al. 2009)</td>
</tr>
</tbody>
</table>
## Personalization

<table>
<thead>
<tr>
<th>Personalization</th>
<th>Treble CLEF User Communities Workshop (Gonzalo et al. 2009)</th>
</tr>
</thead>
<tbody>
<tr>
<td>There must be some support to specify language skills and translation preferences in the user profile.</td>
<td></td>
</tr>
</tbody>
</table>
## Other MLIA tasks

<table>
<thead>
<tr>
<th>Image Retrieval</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Combine text-based with content-based facilities.</td>
<td>(Cheng et al., 2005)</td>
</tr>
<tr>
<td>Advanced organization features can be appreciated by users.</td>
<td>(Petrelli &amp; Clough, 2006)</td>
</tr>
<tr>
<td>Cross-language search should not be offered by default.</td>
<td>(Artiles et al., 2007)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Question Answering</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>When user interaction is possible, simple systems may suffice for CL-QA tasks.</td>
<td>(López-Ostenero et al., 2005b); Peinado et al. 2006; López-Ostenero et al., 2008)</td>
</tr>
<tr>
<td>Users need more context than in a monolingual setting to assess potential answers.</td>
<td>(He et al., 2005; Figuerola et al., 2005; López-Ostenero et al., 2005b; Navarro et al., 2005)</td>
</tr>
<tr>
<td>If feasible, use monolingual IR over a translated document collection as backbone.</td>
<td>(Gonzalo &amp; Oard, 2005; López-Ostenero et al., 2005b; López-Ostenero et al., 2008)</td>
</tr>
</tbody>
</table>
Query Translation & Refinement

Query: European campaigns racism ethnic religious immigrants

Query Terms: European campaigns racism ethnic religious immigrants

Dictionaries: German-One KWIC

- gläubig: believer, believers, creditor, more religious, most religious
- religiös: religiously
- religiös: religiously
- religiös:
- fromm: godly, pious, piously, piously, godlier
- gewissenhaft: sedulous, precise, conscientious, faithful, conscientiousness
Is there CLIR?

Results 1 - 10 of about 372,000 for taipei museum of fine arts. (0.24 seconds)

Tāi-bēi Shīlǐ Měiānguăn, 1983, 24 December
Located in the Beishan Forest Park, occupying
six thousand square meters. The museum has
three main exhibition halls: The Western Art
Academy, The Modern Art Museum, and The
Taiwan Art Museum.

Taipei Fine Arts Museum - Wikipedia, the free encyclopedia
The Taipei Fine Arts Museum (traditional Chinese: 台北市立美術館; pinyin: Táiběi Shīlǐ Měiānguăn; POJ: Tai-pak Chhi-lip Bi-sut-koan) is a museum in Taipei, ...

MOCATaipei
Art Taipei 2009 Year Project - Art & Environment - Human as Art 2009/8/28 ~ 2009/9/1 ...
Museum is closed from Jan 19, 2009 to Feb 5, 2009. ...

TIT Museums: Taipei Fine Arts Museum
Modern art museum specializing in artists from Taiwan, but also includes an international
collection of art.
Is there CLIR?

concerned with an “interchange of time and space?”

Ironically, the wholesale manifestation of Pop would occur in the East with the metamorphosis of comic-book imagery and cheesy science-fiction into manga, anime and tokusatsu. “Desire and Consumption: Kaiyodo and Otaku Culture” is a sprawling nightmarish vision of world domination by Japanese toy manufacturers and savvy illustrators currently on view at the Taipei Fine Arts Museum. Ranging from the innocent shokugan figurines (packaged with food products targeted at children) to the R-rated bishoujo dolls (beautiful young women of manga and anime), this exhibit sheds little light on Pop as theory, extolling instead the skills and business acumen of otaku culture’s various artisans, including the redoubtable BOME and the ubiquitous Takashi Murakami.

The current exhibition of contemporary painters at the Macau Museum of Art exemplifies the disembodiment of theory in contemporary Asian art. Nearly all of the work in “Macau Contemporary Paintings” concerns with “subjective

http://www.tftamuseum/index.aspx

CREATIVE COMMONS LICENSE

All text by Mark Cameron Boyd is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 License. Based on original posts by MCB at theorynow.blogspot.com.
Example approach: using NPs

- abortion
- aborto
- abortion issue
- issue
- tema
- número
- asunto
- edición
- emisión
- abortion issue
- tema del aborto

Corpus evidence:
- tema del aborto
- asunto del aborto
- asuntos como el aborto
- asuntos del aborto
- temas como el aborto
- asunto aborto
NP alignment using comparable corpora

**SPANISH**

acuerdo de libre comercio
acuerdos de libre comercio
acuerdo libre comercio
acuerdo de libre cambio
acuerdos de libre cambio
convenio de libre comercio
convenios de libre comercio
compromiso de libre comercio
...

**ENGLISH**

free trade agreement
free trade accord
free trade pact
free trade beyond the pact
free trade pacts
free trade agreements
free trade arrangements

Fig. 1. Example of sets of noun phrases equivalent under translation.
Greedy translation algorithm (I)

### Iteration 1

<table>
<thead>
<tr>
<th>(C) (set of candidate phrases)</th>
</tr>
</thead>
<tbody>
<tr>
<td>variety of diseases (\rightarrow) tipo de enfermedades (35)</td>
</tr>
<tr>
<td>wide variety (\rightarrow) amplia variedad (29)</td>
</tr>
<tr>
<td>advances in treatment (\rightarrow) avances en el tratamiento (15)</td>
</tr>
<tr>
<td>treatment of a wide (\rightarrow) trato amplio (1)</td>
</tr>
</tbody>
</table>

**Most frequent candidate:** variety of diseases

**Overlapping phrases:**
- wide variety
- variety of diseases

**Partial translation:**

advances in treatment of a wide variety of diseases

\[\text{tipo de enfermedades}\]
Greedy Translation algorithm (II)

**Iteration 2**

<table>
<thead>
<tr>
<th>$\mathcal{C}$ (set of candidate phrases)</th>
</tr>
</thead>
<tbody>
<tr>
<td>advances in treatment $\rightarrow$ avances en el tratamiento (15)</td>
</tr>
<tr>
<td>treatment of a wide $\rightarrow$ trato amplio (1)</td>
</tr>
</tbody>
</table>

**Most frequent candidate:** advances in treatment treatment of a wide

**Overlapping phrases:**

**Partial translation:**

- advances in treatment of a wide variety of diseases
  - avances en el tratamiento
  - tipo de enfermedades
# Greedy Translation Algorithm (III)

## Iteration 3

<table>
<thead>
<tr>
<th>$C$ (set of candidate phrases)</th>
</tr>
</thead>
</table>

Remaining terms: wide

Contextual translations? wide $\rightarrow$ ampio
(from *wide variety*)

**Final translation:**

- advances in treatment of a wide variety of diseases
- avances en el tratamiento amplio tipo de enfermedades
Noun-Phrase Cross-Language Summary

Original document

WORLD CUP SOCCER ’94 / THE FIRST ROUND; SPOTLIGHT; NOT AGAINST BRAZIL

Reuters news service sent a picture of Carlos Alberto Torres, the captain of Brazil’s 1970 World Cup championship team, talking with Lothar Matthaeus, captain of Germany’s team, at a recent practice at Southern Methodist University. The caption information included with the photo identified Torres as a German fan. ELLIOTT ALMOND

Systran translation

FÚTBOL ’94 / EL PRIMER REDONDO DE LA TAZA DEL MUNDO; PROYECTOR; NO CONTRA EL BRASIL

El servicio de noticias de Reuters envió un cuadro de Carlos Alberto Torres, el capitán de Equipo 1970 del campeonato de la taza del mundo del Brasil, hablando con Lothar Matthaeus, capitán del equipo de Alemania, en una práctica reciente en Methodist meridional Universidad. La información del subtítulo incluida con la foto identificó a Torres como ventilador alemán. ALMENDRA DE ELLIOTT

Cross-lingual pseudo-summary

- copa del mundo de fútbol
- primera ronda
- reuters news service
- ver carlos
- campeón de la copa del mundo equipo
- lothar matthaeus
- práctica del reciente universidad metodista del sur
- pie información
- torres aficionados alemanes
- elliott almendra
MT vs NPs: results