Federated Search: Breaking Down the Language Barrier

Abe Lederman, President and CTO
Deep Web Technologies, Inc.
2010 NFAIS Annual Conference
March 1, 2010
About Deep Web Technologies...

- Founded by Abe Lederman, a co-founder of Verity, 2002
- Pioneered federated search technology
- Over $3M in R&D
- Production applications since 1999
- Based in Santa Fe, New Mexico
- 22 person company with strong executive team
Importance of Multilingual Search

- Increases the value of research output by making it available to a wider audience
- Makes available research from China, Japan, Russia, and other countries prolific in science publication
- Greatly broadens the scope of patent research
Importance of Multilingual Search (cont.)

• Exposes English speakers to diverse perspectives from researchers in foreign countries
English Isn’t the Only Language that Matters

Thomson Reuters Research Reveals...

- China’s research output far outpacing the rest of the world
- China surpassed Japan, the UK and Germany in 2006 and now stands second only to the USA
- At this pace, China will overtake the USA within the next decade
- Brazil's share of research output is growing rapidly
Babel Fish Popularized Machine Translation on the Web

- The first European language translation service for web content
- Launched 12/9/97 by DEC’s Alta Vista and SYSTRAN S.A.
- Babel Fish, in "The Hitchhiker's Guide to the Galaxy", is a fish you stick in your ear that allows humans to speak and understand any language
- When released, Babel Fish understood five European languages: French, German, Italian, Portuguese and Spanish
Babel Fish Popularized Machine Translation (cont.)

- SYSTRAN, founded in 1968, leveraged the results of 20 years of military-industrial research
Fun Facts About Machine Translation

• In 1954, the Georgetown-IBM experiment, involved fully automatic translation of more than 60 Russian sentences into English and ushered in the era of significant funding for machine translation.

• The authors of the Georgetown experiment claimed that within three or five years, machine translation would be a solved problem.
Fun Facts About Machine Translation (cont.)

- In the 17th century, philosophers Leibniz and Descartes proposed codes to relate words between languages.
- The first patents for "translating machines" were applied for in the mid 1930s.
- One patent, issued in 1933, was for a storage device using paper tape to find the equivalent of any word in a foreign language.
Approaches to Machine Translation

Rule-based Machine Translation:

- Requires extensive lexicons with morphological, syntactic, and semantic information, and large sets of rules.
- Users can improve the out-of-the-box translation quality by adding their terminology into the translation process.
Approaches to Machine Translation (cont.)

Statistical Machine Translation

- The most widely studied approach to machine translation
- Utilizes statistical translation models whose parameters stem from the analysis of monolingual and bilingual corpora
Approaches to Machine Translation (cont.)

Statistical Machine Translation (cont.)

- Building statistical translation models is a quick process, but the technology relies heavily on existing multilingual corpora
- A minimum of 2 million words for a specific domain and even more for general language are required
Hybrid Machine Translation

- Leverages strengths of rule-based and statistical approaches
- Rules are used to pre-process data in an attempt to better guide the statistical engine
- Rules are also used to post-process the statistical output to perform functions such as normalization
Major Issues with Machine Translation

• Disambiguation - distinguishing between different meanings of a word ("bridging the gap" vs. "dental bridge" vs. "bridge loan" vs. "suspension bridge")

• Harder disambiguation when the text itself is ambiguous
Major Issues with Machine Translation (cont.)

- Idioms - words cannot be translated literally, especially between languages: "hear" vs. "Hear, Hear!"
- Morphology - different word orders
- Words not in the translator's vocabulary
- Translating science has fewer issues
Multilingual Federated Search: State of the Art

- Results merging strategy: Si, Callan, and Others; 2008
- Research into scalable searching of heterogeneous multilingual collections: Powell and Fox; 1998
- Cross-Language Evaluation Forum (CLEF) promotes R&D in multilingual information access
How Multilingual Federated Search Works

1. User enters query in their native language
2. Explorit translator engine translates the query into the right language for each source
3. Explorit submits query to each source
4. Each source returns results in the source’s native language
How Multilingual Federated Search Works (cont.)

5. Explorit translator engine translates the results summaries (title, snippet) into the user’s native language

6. Results summaries from different sources are aggregated

7. Results summaries are ranked

8. Results summaries are displayed to the user
How Multilingual Federated Search Works (cont.)

Foreign language search engines

- German
- Chinese
- Russian

Query to be translated for each source

Query in user’s language

Translator

Results in user’s language

EXPLORIT

Results in source’s language

Ranked results in user’s language
Players in the Machine Translation Space

- One of the oldest machine translation companies, founded in 1968
- Uses hybrid machine translation technology it developed
- Has done extensive work for the US Department of Defense and the European Commission
Players in the Machine Translation Space

- Founded in 2002
- Uses statistical techniques from cryptography
- Applies machine learning algorithms that automatically acquire statistical models from existing parallel collections of human translations
Players in the Machine Translation Space (cont.)

- Uses its own translation software, used SYSTRAN until circa 2007
- Based on statistical machine translation
- Google built a 6-language corpus of 20 billion words' worth of human translations from a large set of UN documents, which are normally available in the 6 UN languages
Players in the Machine Translation Space (cont.)

- Powered by Microsoft Translation
- Based on statistical machine translation
- Once used SYSTRAN, now using system developed by Microsoft Research
WorldWideScience.org is an Excellent Candidate for Multilingual Search

- A global gateway to international science databases and portals
- All content is from national governments or vetted by national governments
- Developed and maintained by the DOE Office of Scientific and Technical Information, OSTI
- One-stop searching
- Will include databases from China, Japan, Korea, Germany, and other non-English countries
Milestones in the History of WorldWideScience.org

Jan. 21, 2007
Global Science Gateway Agreement Signed in London

June 22, 2007
WorldWideScience.org Launched

June 12, 2008
WorldWideScience.org Agreement signed in Korea – formalizes commitment to sustain and grow the service

January 8, 2008
India added to WorldWideScience.org

June 22, 2008
People’s Republic of China joins WorldWideScience.org Alliance

October 15, 2008
People’s Republic of China joins WorldWideScience.org Alliance

WorldWideScience.org to Debut Multilingual Searching

- Deep Web Technologies has partnered with OSTI to introduce multilingual searching to WorldWideScience.org
- Free service to be launched in June
- Launch will be at the International Council for Scientific and Technical Information (ICSTI) meeting in Helsinki in June of this year
- ICSTI oversees the WorldWideScience.org Alliance
1. **Physico-chemical processes of polishing optical glass**

   Original Title: Физико-химические процессы полирования оптического стекла


   The Russian Union Catalog of Scientific Literature

2. **Evaluation of mechanical properties in polyurethane polisher is used as abrasives in the polishing process of optical glass**

   Original Title: Evaluation des caracteristiques mecaniques du polissoir en polyurethane utilise comme porte abrasifs durant le processus du polissage du verre optique


   Article@INIST (France)

3. **Eighth China International Optoelectronic Fair (CIOE) - Antwerp Anwerp Electronic Ansett long Chengdu Electronic Science and Technology**

   Original Title: 第八届中国国际光电博览会 (CIOE) - 安特电子安捷伦科技成都电子

   Colored / colorless optical glass, special optical glass, ... optical processing testing equipment: polishing machine, lens molding equipment, electric cast-type...

   Google China

4. **Automated optical Inspection system - Optical Design - Lens Design - Laser Design - Film Design - Lighting design ...**

   Original Title: 自动光学检测系统 - 光学设计 - 镜头设计 - 激光设计 - 薄膜设计 - 照明设计...

   2007 Nian 3 Ri Yue 4 ... In many cases, process engineers ... optical glass lens molding technology, polishing Common Defects Causes and methods to overcome...

   Google China

5. **Why do some glass polishing powder polishing does not shine? - Has been answered - questions and answers horizon**

   Original Title: 为什么有些玻璃抛光粉不亮呢？ - 已回答 - 天涯问答

   Rare earth polishing powder because of its unique chemical-mechanical action principle and the ... in the process of asymmetric war ... activator LaOB: Or (blue), to enhance...
1. **AGUA CALIENTE Y CALORÍFICACIÓN SOLAR**
   **Original Title:** L'ÉNERGIE SOLAIRE POUR LE CHAUFFAGE ET L'EAU CHAURE
   **Article:** ENSIEF (France)

2. **Agua caliente**
   **Title:**
   **Author:**
   **Match:** All Field(s)
   **Date Range:** Pick Year to Pick Year
   **Translate from/to:** Spanish

3. **Agua caliente**
   **Original Title:**
   **PubMed**

4. **Agua caliente**
   **Original Title:**
   **Bibliography:**
   **The Russian Union Catalog of Scientific Literature**

5. **Cálculo de calentadores de agua caliente**
   **Original Title:**
   **Catalogue of the TIB, German National Library of Science & Technology (TIBKat)**
Agua caliente de fría. El disociativas recomendaría...

Original Title: *Hot Water from Cold. The Dissociatives Recommend.*

*Thomas, R. D.; Zhaunerchyk, V.; Hel*  
*The journal of physical chemistry. A*  
*2010-02-19 PubMed*

---

Agua caliente de baterías de tanques metal-manual de Alc. 1/52/7: n° de utv.Roskommunenergo. WA - communes.HOZ - WA RSFSR 11.11.86.

Original Title: *Типовая инструкция по эксплуатации металлических баков-аккумуляторов горячей воды: Утв. Роскоммунаэнерго М-ва жил-коммун. хоз-ва РСФСР 11.11.86*  
*The Russian Union Catalog of Scientific Literature*

---

Calculo de calentadores de agua caliente

Original Title: *Berechnung von Warmwasser-Heizungen*  
*Böhmle, Hubert*  
*1982-01-01*  
*Catalogue of the TIB, German National Library of Science & Technology (TIBKat)*
References

An effective and efficient results merging strategy for multilingual information retrieval in federated search environments
http://portal.acm.org/citation.cfm?id=1331574

Babel Fish

China's Research Output More than Doubled Since 2004, Thomson Reuters Study Reveals

Comparison of machine translation applications

Cross Language Evaluation Forum
http://www.clef-campaign.org/

Deep Web Technologies Developing Multilingual Translator for Federated Search
http://www.ereleases.com/pr/deep-web-technologies-developing-multilingual-translator-federated-search-25166

Deep Web Technologies to unveil multilingual federated search in June
References (cont.)

Deep Web Implements the Multilingual Search that Google Imagines

History of machine translation

Machine translation
http://en.wikipedia.org/wiki/Machine_translation

Multilingual Federated Searching Across Heterogeneous Collections
http://www.dlib.org/dlib/september98/powell/09powell.html

SYSTRAN: What is Machine Translation?

Thomson Reuter Global Research Report Series
http://researchanalytics.thomsonreuters.com/grr/

WorldWideScience.org News/Press Releases
http://worldwidescience.org/news.html
Thank you!

Abe Lederman
abe@deepwebtech.com
Online Presentation:
http://deepwebtech.com/talks/NFAIS.pdf