Multilingualism on the Internet: Putting the World in the World-Wide Web

Arle Lommel
Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)
SOME BACKGROUND
This works at the data level: you can access the Internet from (almost) anywhere in the world.
The Reality: Multiple Webs

Top Ten Languages on the Internet

English
Spanish
French
Russian
Arabic
German
Japanese
Korean
Chinese

Based on
Change

• Since 2000:
  – Chinese grew ~1500%
  – Arabic grew ~2500%
  – Russian grew ~1800%
  – Top ten languages grew 421%
  – Rest of the world grew 589%

Based on
Alexa Rankings in Germany

- Google.de
- Facebook.com
- Google.com
- Youtube.com
- Ebay.de
- Amazon.de
- Wikipedia.org
- Spiegel.de
- Bild.de
- Yahoo.com
Alexa Rankings in China

- Baidu.com
- QQ.com
- Taovao.com
- Sina.com.cn
- Google.com.hk
- 163.com
- Weibo.com
- Sohu.com
- Google.com
- Ifeng.com
Why the Difference?

- Culture: Different sites are adapted to different cultures
- Language: You can’t access what you can’t read
Culture

- Difference is good…
- You can (usually) deal with cultural difference
- It enriches our experience
- Great if we could access Chinese sites to learn…
Language is the Key Issue

• Language both enables and prevents cultural contact

• Solutions?
  – We all learn the same language
  – We all learn every language
  – Translation
Two Translation Approaches

• Push: The content creator supplies the translation: E.g., translation company does the work

• Pull: The content consumer gets the translation: E.g., Google Translate
Why Not Human Translation for Everything?

- Expensive (€0,25/word is common): a 5000-word document can cost €1.250 for one language
- Not enough translators
- Not suitable for on-demand needs
- Not all languages are available
Why Not Google Translate for Everything?

- Quality is a major issue
- Lack of integration: not easy to access
- No social interaction
- Relies on human translation
Intelligence Creates Quality

• To fix quality (for humans or machines), you need intelligence…

• Intelligence allows you to deduce things you would not otherwise be able to understand

• Should “Different Types of Windows Are Very Useful” be?
  – Verschiedene Arten von Windows sind sehr nützlich
  – Verschiedene Arten von Fenster sind sehr nützlich
Humans Are...

• Very good at:
  – deduction
  – dealing with (implicit) context
  – adding intelligence
  – dealing with exceptions

• Very bad at:
  – doing things quickly
  – doing the same thing over and over
Machines Are...

• Very good at:
  – doing things quickly
  – doing the same thing over and over

• Very bad at:
  – deduction
  – dealing with (implicit) context
  – adding intelligence
  – dealing with exceptions
If Machines Cannot Add Intelligence...

• Then we need to make data intelligent
• Apply what humans do well to enable machines to do what they do well
Example

• An alcoholic beverage slogan:
• “Give the Gift of Mist”
• Do we translate this or not?
Examples

• The “translate” attribute (already in HTML5):
  – `<p>He bought a `<span translate="no">Super Soaker™</span>` water gun at `<span translate="no">Joe’s Toy Emporium</span>`</p>`

• “Domain”:
  – Windows = Windows (IT)
  – Windows = Fenster (Construction)
The MultilingualWeb-LT Project

• Seeks to…
  – Supply **metadata** to support the translation of Web and “deep Web” content
  – Simplify translation (and lower costs)
  – Improve quality of translation through intelligent data
  – Create easy ways for content creators to improve their source text
The MultilingualWeb-LT Project

• Administered through the World Wide Web Consortium
• Funded by the European Commission
• Run through the German Research Centre for Artificial Intelligence
• Representatives of industry, research, academia
• Focus on concrete deliverables
Proposed Scope
Intended Outcome

• Lower barriers to information access
• Lower cost of translation (for human translation)
• Increase quality (for all translation)
• Enable linking of data across languages (improve quality for all)
So What Do We Need?

- Better dynamic translation tools: e.g., on online forums
- Better quality
- Easier access to translation
- Planning for translation
- Support