PangeaMT – putting open standards to work... well

Manuel Herranz – PangeaMT - Pangeanic

www.pangea.com.mt
Unmanageable amounts of data? *The data deluge*

- As of May 2009: 487 Billion gigabytes or 1,000,000,000 * 487,000,000,000 = $4.87 \times 10^{20}$

- Estimates
  - Up 50% a year (Oracle)
  - Doubles every 11 hours (IBM)

- Language translation as a job becoming unmanageable. Increasing demands, increasing volumes, shorter deadlines. Human production is not sufficient.
Short history

- Pangeanic: LSP. Major clients in Asia, European localization, increasing number of languages and volumes

- Need to produce faster, cheaper, quality
  - Experimenting with some RB systems
  - TAUS & TDA founding members (M's of words!)

- Partnering with Valencia's Computer Science Institute (R&D and EU projects: Casacuberta, Och, Vidal, Koehn)
Short history

- **CHALLENGE:** Turn academic development (Moses) into a *commercial application*.

- Limitations: plain text (txt), language model building (first), no reordering, no updating features (always re-start), data availability, Linux-based (server). You need computational linguists (programmers), not translators, to operate it.

- Partnering with Valencia's Computer Science Institute PangeMatic (v1) was developed and then PangeaMT 2009 (web-based)
Short history

- OBJECTIVES:

1. To provide High Q MT for Post-Editing and save time and cost. No Google-type broad TR but domain-specific, user-centric.

2. To use only community-based Open standards → Oasis / ISO: xlfiff / tmx, xml). NO proprietary formats (technology independence) so clients are not “locked” in to buying and updating expensive software.

3. To automate as many processes as possible.
Short history - Implementations

Plus many other internal engines for...

* Large Japanese Car manufacturing firm
* Electronics firms
* Technical / Engineering
How PangeaMT works

Use Open Standards Browser: Mozilla, Safari

About PangeaMT

PangeaMT builds on SMT technologies that have been tested and applied by Pangeanic for its clients.

The company is privately owned and offers both consulting and systems integration. We train Statistical Machine Translation engines and customize them to your company’s terminology and expressions. We make them fit in your current system so you do not need to purchase additional software.

PangeaMT is a European leader in the implementation of independent, open source statistical
How PangeaMT works

pangea.com.MT
The Statistical MT Customization Company

Plain Text (without tags)

1. Select File to translate
2. Select language pair

3. Select type of engine

- TMX / XLIFF / TXT File
- Language Pair: Spanish > English
- Engine:
  - Select legal 2010
  - LSM 2010
  - MEB 2009
  - McLEHM Specific
  - McLEHM Extended

Translation

PangeaMT – putting open standards to work... well
How PangeaMT works

Users get an email with the translation minutes later
Post-editing

In any case, there is a wide variation in the balance sheet, profitability indicators and bad debts of these organisations.

Thus, in the first half of 2010 there have been intensive process of concentration which, in some cases, have been formalized through the

SET MT MATCH PENALTY

PangeaMT – putting open standards to work... well
Future Work

- “on the fly” MT training (minutes, not manually) – April 2011 !!

- pick and match sets of data: “extreme customization” – April 2011 !!

- objective stats for post-editors (calculate effort)

- confidence scores for users (→ translators or readers) with CAT integration (web-based / desktop)

- Web samples
Thank you!

QUESTIONS?

mherranz@pangea.com.mt