Machine Translation for Translators

Ruben de la Fuente

American Translators Association
About Me

• 4-year degree in translation
• Worked as translator for 10+ years
• Only working full time in MT for the past year
Do You Think MT Will Be Part of Your Toolbox in a Future?

Poll
Why should you concern yourself with Machine Translation?

- Produce the **same quality**
- Only **faster**
- **Make more revenue**
Does MT Make Sense For You?

- Volumes are high
- Texts are consistent in terms of terminology and style
- Domain does not matter so much if texts are consistent
Customize to suit your needs

- Customize upfront
- Update regularly

MT engines are not finished products, but on-going work
Keys to evaluate MT

Use hard facts and not impressions:

• Run an automatic QA tool (Xbench): flags obvious errors (terminology, capitalization, punctuation)

• See how much you can edit in one hour

• Calculate edit distance and generate change reports (SymEval)
Checklist for MT Tool Selection

• Linguistic quality
• Language pairs
• Ease of integration
• Customizability
• Confidentiality
Rule-based MT

- Dictionaries+Translation rules
- Output is grammatically correct and predictable, but maybe not very natural
- More costly and take longer to develop
- More user-friendly
ProMT

• Demo fully functional for 30 days
• Available languages: EN<>FR, IT, DE, ES, RU, PT
• Download Language Service Provider 9.5
Profiles

- Dictionary hierarchy
- Translation rules
- Translation memories and thresholds
Translating with ProMT

Demo
Always Save Output to TMX
Statistical Machine Translation

- Figure out word/phrase alignment by statistical analysis of bilingual corpora
- Output is **not predictable**, but can be more fluent than RbMT
- **Cheaper** and **quicker** to develop
- Require more technical skills. Very **user-unfriendly** for now.
Caveat

Garbage in, garbage out principle was never truer

Make sure TMs are in good shape
Do-Moses-Yourself (DoMY)

**Graphs:** import-tmx, clean-LM/TM, build LM/TM, train, translate.

**Ini files:** configuration (language pairs, paths for input and output).

**Folder structure:** hierarchy and stages
Running DoMY

- **Command line:**

  ```
  C:\Users\rubo>domy import tnx
  ```

- **Ini for configuration**

  ```ini
  [USER]
  source= REQUIRED
  target= REQUIRED
  superdomains=* REQUIRED
  domain=*
  subdomain=*
  filespec=*.*
  buildname=
  lowmem=True
  removegram=False
  removedupe=False
  ```

  default
  Flag to process language model data in smaller pieces to run with lower memory
  Remove 1-gram phrases from consolidated corpus file
  Remove duplicate phrases from the consolidated corpus
DoMY Output

• **Translation Model**: table containing source and target phrases, together with a probability score.

• **Language Model**: monolingual corpus the system refers to produce more fluent output (e.g. reordering)
Language and Translation Models

- **LM** (fluency)
  - iARPA

- **TM** (equivalences)
  - example.com ||| http://www.example.com ||| 1 0.040007 6.18417e-06 2.718 |||| 1
  - a message ||| un mensaje ||| 1 0.731223 1 0.492059 2.718 |||| 1
  - a ||| un ||| 1 0.132346 1 0.333333 0.877566 2.718 |||| 1 3
  - a ||| un ||| 1 0.130104 0.333333 0.007048 2.718 |||| 1 3
  - a ||| un ||| 0.0359444 0.333333 0.333333 0.0359444 2.718 |||| 1 3
  - according ||| de acuerdo con ||| 1 0.014361 0.0031796 0.718 |||| 1 1
  - all ||| los ||| 0.142957 0.0456125 0.333333 0.0456125 2.718 |||| 1 7
  - all ||| los ||| 0.397539 0.0456125 0.333333 0.273077 2.718 |||| 1 3
  - and a complaint message to ||| y un mensaje de reclamación para ||| 0.8 0.0359444 0.0031796 2.718 |||| 1 2 1
  - and a complaint message to ||| y un mensaje de reclamación para ||| 0.8 0.0359444 0.0031796 2.718 |||| 1 2 1
  - and display message to ||| y muestre mensajes para ||| 0.5 0.00004173 1 0.00004173 2.718 |||| 1 2 1
  - and display message to ||| y muestre mensajes para ||| 0.5 0.00004173 1 0.00004173 2.718 |||| 1 2 1
  - and display ||| muestre ||| 0.5 0.0359444 0.0031796 2.718 |||| 1 2 1
  - and the template of the ||| y la plantilla del ||| 1 0.5 0.003994799 1 0.0546125 2.718 |||| 1 2 1
  - and the ||| y la ||| 0.5 0.0179457 2.718 |||| 1 2 1
  - and the ||| y la ||| 0.333333 0.333333 0.333333 2.718 |||| 1 3 2
  - and the ||| y la ||| 0.8 0.34001 0.142957 0.142957 2.718 |||| 1 3 7
  - and the ||| y la ||| 0.0359444 0.0359444 0.0359444 2.718 |||| 1 2 1

American Translators Association
M4Loc

Set of scripts to integrate Moses with L10N tools:

• **Xliff to Moses**: remove mark up
• **Moses to xliff**: reinsert mark up
• **Also include Adobe Moses toolset**
Where To Get Corpora

• Opus (ECB, EMEA, OpenOffice)
• Acquis Communautaire
• Europarl
• Multilingual websites: Bitextor
Tips for Efficient Postediting

• Embrace the keyboard
• Run automatic QA tool
• Fix repetitive issues with global search and replace (powered with wildcards and regular expressions)
• Store S&R operations in macros
RegEx Basics

- Charsets: [a-z] [0-9]
- Non-printable: \n, \r, \t
- Anchors: ^, $, \b
- Alternation: |
- Grouping: ()
Search and Replace Macro

Demo
Do You Still Think it Takes Less Time to Translate from Scratch?

Poll
Questions?
Speak now…
Or reach me at:
www.facebook.com/xlation
www.wordbonds.es
@rubendelafuente
http://www.linkedin.com/in/rubendelafuente