Let's MT! — A Platform for Sharing SMT Training Data

Jörg Tiedemann, Per Weijnitz

jorg.tiedemann@lingfil.uu.se, per.weijnitz@convertus.se
Department of Linguistics and Philology
Uppsala University

October 2010

The Project

Let's MT!

http://www.letsmt.eu/

- Funded under: The Information and Communication Technologies Policy Support Programme (ICT PSP)
- Theme 5 - Multilingual web, grant agreement no 250456.

LetsMT! Goals

Develop an online collaborative platform for **data sharing** and **MT building**

- based on existing open SMT technologies
- address private users, academic users, commercial users
- support for under-resourced languages
- support for domain/user-specific collections

Project Partners

- **Tilde** SIA, Riga, Latvia
- **University of Edinburgh**, Human Communication Research Centre (HCRC), Edinburgh, UK
- **University of Zagreb**, Faculty of Humanities and Social Sciences, Department of Linguistics, Zagreb, Croatia
- **University of Copenhagen**, Centre for Language Technology, Copenhagen, Danmark
- **Uppsala University**, Department of Linguistics and Philology, Uppsala, Sweden
- **SemLab/Zoorobotics** BV, Alphen a/d Rijn, Netherlands
- **Moravia**, Brno, Czech Republic

Essential Features

- resource repository with SMT training data
- upload facilities & data management
- data sharing & data security
- user-specific training of SMT models
- on-line translation service
- integration in web browsers and CAT tools

Development

- build facilities for data storing and sharing
  - aligned parallel data (TMX, XLIFF, ...?)
  - non-aligned parallel data (PDF, DOC, TXT, ...?)
    → integrate automatic sentence alignment
    → allow human control (cleaning, rating, ...)
  - monolingual data (various formats)
  - browsing, selecting, permission control

Development

- build facilities for data storing and sharing
  - aligned parallel data (TMX, XLIFF, ...?)
  - non-aligned parallel data (PDF, DOC, TXT, ...?)
    → integrate automatic sentence alignment
    → allow human control (cleaning, rating, ...)
  - monolingual data (various formats)
  - browsing, selecting, permission control
- fill data repository with available data sets
  - available parallel corpora (all partners)
  - available monolingual corpora (all partners)
  - language-specific tools (tokenizers, segmenters, ...?)

Development

- integrate SMT training pipe line
    - standard Moses/Giza++ & friends
    - grid engine/cloud solutions
    - simplicity first → address non-technical users
    - allow parameter adjustments → advanced users

Development

- integrate SMT training pipe line
  - standard Moses/Giza++ & friends
  - grid engine/cloud solutions
  - simplicity first → address non-technical users
  - allow parameter adjustments → advanced users
- provide translation services for a selected number of languages
  - provide baseline systems
  - run a number of engines (to be decided)

The LetsMT Data Repository

General framework:

- Webservice API (REST)
- off-line data processing (validation, conversion, ...)
- backend: version-controlled file system

The LetsMT Data Repository

General framework:

- Webservice API (REST)
- off-line data processing (validation, conversion, ...)
- backend: version-controlled file system

Sharing via branching:

- authorized users can create branches of existing resources
- branching secures data integrity & storage efficiency
  - space-efficient (diff's only)
  - each branch can be modified independent of others
- permissions: private, shared, public

Internal Storage Format

- standalone XML for corpus data
- external sentence alignment

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE cesAlign PUBLIC "-//CES//DTD XML cesAlign//EN" "">
<cesAlign version="1.0"><linkList><linkGrp targType="s"
  fromDoc="https://letsmt.eu/storage/Europarl/xml/eng/ep-00-01-17.xml"
    toDoc="https://letsmt.eu/storage/Europarl/xml/fre/ep-00-01-17.xml">
      <link xtargets="1;1" />
      <link xtargets="2;2" />
      <link xtargets="3;3 4" />
```

Internal Storage Format

- standalone XML for corpus data
- external sentence alignment

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE cesAlign PUBLIC "-//CES//DTD XML cesAlign//EN" "">
<cesAlign version="1.0"><linkList><linkGrp targType="s"
  fromDoc="https://letsmt.eu/storage/Europarl/xml/eng/ep-00-01-17.xml"
    toDoc="https://letsmt.eu/storage/Europarl/xml/fre/ep-00-01-17.xml">
      <link xtargets="1;1" />
      <link xtargets="2;2" />
      <link xtargets="3;3 4" />
```

Advantages:

- can link documents to multiple translations without copying
- can handle sentence alignment variants
- support manual alignment manipulation without data manipulation
- simple corpus selection (several corpora, sub-corpora, 1:1 only, ...)

Training User-Tailored SMT models

Important goal: Support building user-specific SMT models!

- Let'sMT user may select training data they need
- Let'sMT builds standard phrase-based SMT based on user selection

How much can we gain?

Domain-specific Translations: EMEA

Experiments with EMEA
(from http://www.let.rug.nl/tiedeman/OPUS/)

	English	Swedish
sentences	898,359	898,359
tokens	11,567,182	10,967,600
unique sentence pairs		
sentences	298,974	298,974
tokens	4,961,225	4,747,807

→ Highly repetitive texts with very consistent terminology!

Domain-specific Translations: EMEA

Standard setup with Moses & friends:

- data sets (from unique set of sentence pairs):
  - 1000 randomly selected pairs for tuning
  - 1000 randomly selected pairs for testing
  - remaining for training
- language model: 5-gram (SRILM)
- translation model: standard Moses/Giza++ settings
- tuning: standard MERT

Comparison: General-purpose engine "Google translate"

Domain-specific Translations: EMEA

And the results are:

BLEU in %	Google (08/2010)	Moses-EMEA
English-Swedish	50.23	
Swedish-English	46.57	

Domain-specific Translations: EMEA

And the results are:

BLEU in %	Google (08/2010)	Moses-EMEA
English-Swedish	50.23	**59.29**
Swedish-English	46.57	**65.42**

Wow!

Conclusions

- collaborative platform for sharing SMT data
- user-friendly interface to open SMT tools
- customer-specific SMT models
  → **Large performance gains possible!**
- online translation services
- browser widgets & SMT integration in CAT

Let's MT! (... stay tuned)