CNGL
Centre for Next Generation Localisation

OpenMaTrEx: A Free/Open-Source Marker-Driven Example-Based MT System

S. Dandapat, M.L. Forcada, D. Groves, S. Penkale, ***John Tinsley*** and A. Way
School of Computing, Dublin City University, Ireland

IceTAL 2010, August 18th, Reykjavik, Iceland

NDP National Development Plan 2007 - 2013
sfi science foundation ireland
DCU Dublin City University
UCD Dublin University College Dublin
University of Limerick
Trinity College Dublin

Overview

- Introduction and Background
- Freeing/open-sourcing MaTrEx
- System Overview
- Functionality
  - training
  - translating
- Sample Experiment
- Future Release Cycle
  - New Components
- Acknowledgments

Freeing/open-sourcing MaTrEx

- Collaborative effort to combine a number of individual components
  - identify various translation workflows and provide unified support (still a work-in-progress!)
  - begin together code from researchers personal configurations
  - identifying authors of components and obtaining permission from IP holders

- This gave rise to a number of outcomes
  - identification and fixing of several bugs
  - creation of new FOS marker word files for languages
  - development of improved t-table merging procedure

Intro and Background

- MaTrEx developed based on successful research by the Dublin City University MT group
  - scores of publications based on components, particularly in terms of shared task participations (e.g. WMT 08-10, IWSLT 06-09, ICON 08, NTCIR 10)
- OpenMaTrEx is a free/open-source (FOS) implementation of the basic MaTrEx components
  - wrapper around existing FOS software e.g. Giza++, Moses
  - marker-based EBMT component – Marclator
- Freeing/open-sourcing MaTrEx guarantees **reproducibility** and encourages **collaborative research**

System Overview (1)

OpenMaTrEx is itself a hybrid example-based/statistical MT system containing:

- a marker-word driven **chunker**, with marker word files for a number of languages;
- a collection of chunk **aligners** using a variety of distance algorithms;
- tools for **merging** translation tables;
- two full MT engines
  - a proof-of-concept monotone chunk-based engine (example-based recombiner);
  - a wrapper around the statistical MT engine Moses (also FOS).
- Support is also included for automatic evaluation

System Overview (2)

- Central to the novel EBMT components of OpenMaTrEx (a.k.a. Marclator) is the **marker hypothesis** (Green, 1979)
  - all natural languages are marked for complex syntactic structure at surface form by a closed set of specific lexemes and morphemes: *markers*.

- Marker words may be used to chunk the text:

  [**He** came] [**from the** office] [**to** witness] [**the** chemical process.]

  [Vino] [**del** despacho] [**para** presenciar] [**el** proceso químico.]

- Aligned chunk pairs form *subsentential translation units*:

  [**from the** office] ↔ [**del** despacho]

Functionality - training

- **Marclator ("example-based") mode**
  - source and target sentences chunked using marker words
  - subsentential chunks are aligned (edit distance, with/without jumps)
    - edit distance costs can include external word alignment probabilities
- **MaTrEx mode**
  - run Giza/Moses components for alignment
  - *merge* Moses and Marclator chunks
    - re-estimate phrase translation probabilities
    - (optional) feature indicating origin of translation pair

- MERT for tuning

Functionality - translation

- **Marclator mode**
  - monotone (naïve) decoder
  - input is chunked based on marker words and most probable chunk selected
  - if no translation, back of to Giza++ word translations
- **MaTrEx mode**
  - Moses decoder is run on the merged phrase table
- Other decoders may be plugged in – support included in future releases

Sample Experiment

- 200k sentences selected at random from Spanish-English WMT08 workshop data
- Tuning and testing done of sets provided by WMT08

System	BLEU	NIST	EBMT pairs
Baseline Moses	30.59%	7.517	27.6%
MaTrEx mode	30.42%	7.516	29.5%
MaTrEx mode + feature	30.75%	7.527	33.6%

Future Releases

- Current version 0.9 – Meteor support, large-scale LMs
- Regular releases to include:
  - improved marker files
  - improved installation process and documentation
  - improved running procedures for training and testing
  - freeing/open-sourcing of other successful MaTrEx modules and adding them as *new components*:
    - word-packing (Ma, 2009)
    - source-side context (Haque, et al. 2009)
    - sub-tree alignment (Tinsley and Way, 2010)
- Increase interoperability with other FOS tools
- Create a community of developers and contributors

Acknowledgments (and promotion!)

- Thanks to authors past and present for their collaborative efforts:
  - Steven Armstrong, Pratyush Banjeree, Sandipan Dandapat, Mikel Forcada, Yvette Graham, Declan Groves, Hany Hassan, Yanjun Ma Bart Mellebeek, Jimmy O’Regan, Pavel Pecina, Nicolas Stroppa and Andy Way

http://www.openmatrex.org

Thank you