MT Research in the NCLT and the CNGL

Andy Way

CNGL & NCLT
School of Computing
Dublin City University
Dublin, Ireland

away@computing.dcu.ie

NGL CSET Scientific Committee

23/7/08
Presentation

Overview

Context

Dependencies

Research Tracks

Core Machine Translation Engines

Large-Scale MT Evaluation

Future Developments
Contents of Presentation

- CSET Objectives & Outcomes
- Dependencies
- Research Tracks in the CSET
  - Core Machine Translation Engines
  - Exploiting Synergies between Speech Technology and MT
  - Text Classification and Automatic Labelling
- Large-Scale MT Evaluation
- Future Developments
Presentation

Overview

Context

Dependencies

Research Tracks

Core Machine Translation Engines

Large-Scale MT Evaluation

Future Developments
Facilitating Optimal Multilingual NGL Applications

Objectives

- Undertake research into the tight coupling of speech technology and MT for improved multilingual applications;
- Reliable automatic classification and annotation of data to:
  - enable optimal tuning of MT and spoken language systems in NGL applications;
  - facilitate improved processing of multilingual queries (DCM);
  - significantly improve the workflows of our industrial partners (LOC).
Facilitating Optimal Multilingual NGL Applications

Objectives

- Undertake research into the tight coupling of speech technology and MT for improved multilingual applications;
- Reliable automatic classification and annotation of data to:
  - enable optimal tuning of MT and spoken language systems in NGL applications;
  - facilitate improved processing of multilingual queries (DCM);
  - significantly improve the workflows of our industrial partners (LOC).

Outcomes

- Novel and improved MT and Speech Recognition & Synthesis engines for multilingual NGL applications;
- Novel automatic annotation of digital content for NGL.
Presentation

Overview

Context

Dependencies

Research Tracks

Core Machine Translation Engines

Large-Scale MT Evaluation

Future Developments
Outputs
Presentation

Overview

Context

Dependencies

Research Tracks

Core Machine Translation Engines

Large-Scale MT Evaluation

Future Developments
ILT Basic Research Areas

- Core Machine Translation Engines (Area Leader: Prof. Andy Way, DCU)
- Exploiting Synergies between Speech Technology and MT (Prof. Julie Berndsen, UCD)
- Text Classification and Automatic Labelling (Dr. Carl Vogel, TCD)
Presentation

Overview

Context

Dependencies

Research Tracks

Core Machine Translation Engines

Large-Scale MT Evaluation

Future Developments
Work Packages

- WP1.1: Incorporating Syntax into SMT Systems
- WP1.2: Improved Hybrid Systems
- WP1.3: Scaling More Linguistically Sophisticated Systems
- WP1.4: Probabilistic Transfer
- WP1.5: Tuning MT
- WP1.6: General Alignment Models
- WP1.7: Automatic MT Evaluation
- WP1.8: Controlled Language
- WP1.9: Patients with Limited English
WP1.1: Incorporating Syntax into SMT Systems

- PI: Prof. Andy Way (DCU)
- Collaborators: Dr. Khalil Sima’an, Prof. Antal van den Bosch, Dr. Salim Roukos, Dr. Josep Crego
- Appointments: PostDoc1, Hala Maghout
- Industrial Contributors: IBM, Microsoft, Symantec, Traslán

Areas of Investigation:
- incorporating supertags in translation and target language models (cf. [Hassan et al., 06, 07, 08]);
- incorporating source language features (cf. [Carpuat & Wu, 2007; Stroppa et al., 2007]).

Other Collaborators:
- John Tinsley (ATTEMPT)
- Hany Hassan (Prospect, IBM)
WP1.2: Improved Hybrid Systems

- PI: Prof. Andy Way (DCU)
- Collaborators: Dr. Declan Groves (Traslán), Dr. Haifeng Wang, Dr. Ralf Brown, Prof. Sadao Kurohashi
- Appointments: Dr. Jinhua Du, Tsuyoshi Okita
- Industrial Contributors: IBM, Microsoft, Symantec, DNP
- Areas of Investigation:
  - adding EBMT-style decoders (cf. [Groves, 2007]);
  - incorporate generalized templates (e.g. [Way & Gough, 2003]) into PB-SMT;
  - extend EBMT system with generalized templates based on content words (cf. [Brown, 1999]).
- Other Collaborators:
  - John Tinsley (ATTEMPT)
  - Sudip Naskar (Prospect)
WP1.3: Scaling More Linguistically Sophisticated Systems

- PI: Prof. Andy Way (DCU)
- Collaborators: Dr. Alon Lavie, Dr. Khalil Sima’an
- Appointments: PostDoc1, Sergio Penkale
- Industrial Contributors: IBM, Microsoft, Symantec
- Areas of Investigation:
  - scale tree-to-tree systems to real scenarios (cf. [Hearne & Way, 2006]);
  - improving search-based syntax-driven MT systems (cf. [Lavie et al., 2008]).
- Other Collaborators:
  - John Tinsley (ATTEMPT)
  - Ventzi Zhechev (ATTEMPT)
WP1.4: Probabilistic Transfer

- PI: Prof. Josef van Genabith (DCU)
- Collaborators: Prof. Mikel Forcada
- Appointments: PostDoc2
- Industrial Contributors: IBM, Microsoft, Symantec, DNP
- Areas of Investigation:
  - scaling and improving probabilistic dependency-based transfer systems (cf. [Riezler & Maxwell, 2005])
- Other Collaborators:
  - Yvette Graham
WP1.5: Tuning MT Systems

- PI: Prof. Josef van Genabith, Prof. Andy Way (both DCU)
- Appointments: Dr. Patrik Lambert, Silke Theison
- Industrial Contributors: IBM, Microsoft, Symantec
- Areas of Investigation:
  - Customizing above MT systems to text type and genre of industrial partners’ documentation (UI, Help, Software etc.)
WP1.6: General Alignment Models

- PI: Prof. Andy Way (DCU)
- Collaborators: Dr. Declan Groves (Traslán), Dr. Alon Lavie, Dr. Khalil Sima’an
- Appointments: Dr. Patrik Lambert, Ankit Srivastava
- Industrial Contributors: IBM, Microsoft, Symantec, Traslán
- Areas of Investigation:
  - Improved Models of Word Alignment (cf. [Ma et al., 2007; Lambert & Banchs, 2008]);
  - Sub-sentential Alignment for MT (cf. [Tinsley et al., 2007; Lavie et al., 2008]).
- Other Collaborators:
  - Ventzi Zhechev (ATTEMPT)
  - Yanjun Ma (Prospect)
  - John Tinsley (ATTEMPT)
  - Sylwia Ozdowska (Prospect)
WP1.7: Automatic MT Evaluation

- PI: Prof. Andy Way, Prof. Josef van Genabith (DCU)
- Collaborators: Prof. Hermann Ney, Dr. Alon Lavie
- Appointments: Dr. Jinhua Du, Yifan He
- Industrial Contributors: IBM, Microsoft, Symantec
- Areas of Investigation:
  - Improved non-$n$-gram-based MT Evaluation Methods (cf. [Owczarzak et al., 2008])
WP1.8: Controlled Language

- PI: Dr. Sharon O’Brien, Dr. Dorothy Kenny (both DCU)
- Collaborators: Prof. Andy Way (DCU)
- Appointments: Steven Doherty
- Industrial Contributors: Symantec
- Areas of Investigation:
  - Adapting above MT systems to industrial partner’s controlled language documentation.
WP1.9: Patients with Limited English

- PI: Prof. Harold Somers (DCU)
- Appointments: Dr. Sara Morrissey, PhD1, PhD2
- Industrial Contributors: IBM, Microsoft, Symantec
- Areas of Investigation:
  - Appointment Scheduling (cf. [Somers, 2007])
  - Sign Language Translation (cf. [Morrissey, 2008])
Presentation

Overview

Context

Dependencies

Research Tracks

Core Machine Translation Engines

Large-Scale MT Evaluation

Future Developments
MaTrEx System

Overview
Context
Dependencies
Research Tracks
Core Machine Translation Engines
Large-Scale MT Evaluation
Future Developments
MaTrEx II

- Needs an Overhaul:
  - Not all research improvements incorporated (cf. WMT-08 paper);
  - New *internal* expertise
  - New *external* research findings
- ‘Time Out’ (November?) where all MT developers work on system update
- Think about:
  - Patenting
  - Open-Source Components (cf. Mikel Forcada’s visit)

→ continued improvement in large-scale evaluation campaigns
CSET Languages covered

- French,
- Spanish,
- Romanian,
- Chinese,
- Irish,
- Hindi,
- Arabic,
- Japanese,
- German.
Languages covered

- French,
- Spanish,
- Chinese,
- Irish,
- Hindi,
- Arabic,
- Japanese,
- German,
- Bengali.
Languages covered

- French,
- Spanish,
- Chinese,
- Irish,
- Hindi,
- Arabic,
- Japanese,
- German,
- Bengali.
Languages covered

- French,
- Spanish,
- Chinese,
- Irish,
- Hindi,
- Arabic,
- Japanese,
- German,
- Bengali,
- Polish.
Presentation

Overview

Context

Dependencies

Research Tracks

Core Machine Translation Engines

Large-Scale MT Evaluation

Future Developments
Purpose of this Workshop

- Identify trends, convergences, and any gaps that need filling.
- This will, hopefully, provide strong pointers to the future direction of our research, in the short- to medium-term, at least.
- Research Plans
- Other Research Avenues
  - Indian languages (Hindi, Bengali ... English), e.g. FIRE project
  - etc ...
Other Meetings

Internal

- Students & Postdocs: ongoing
- Students & PI: weekly/fortnightly
- Postdocs & PI: fortnightly
- Groups (e.g. Alignment, SMT): monthly
- Whole Team: two-monthly
Other Meetings

Internal

- Students & Postdocs: ongoing
- Students & PI: weekly/fortnightly
- Postdocs & PI: fortnightly
- Groups (e.g. Alignment, SMT): monthly
- Whole Team: two-monthly

External

- ILT: other academics (TCD, UCD)
- MT: industrial partners
- ILT: industrial partners
- Other CSET R&D strands