CENTRE FOR TRANSLATION STUDIES
at the University of Leeds

Developments in MT
- from now to the near future

Tony Hartley
a.hartley@leeds.ac.uk
Centre for Translation Studies, U Leeds
Graduate School of Education, U Tokyo

UNIVERSITY OF LEEDS

We need a better grasp of MT's imperfections

- We want high quality MT
- The reality is different
- Flaws or 侘寂 ?
- What are the remedies?

This talk visits four main points

- Limitations of main MT architectures
- Growing MT use and availability
- Hybridisation of architectures
- Tomorrow's agenda

Knowledge-driven MT (RMT) tries to render all the source text

Rules abstract away from text surface

- SL & TL morphology, grammar, dictionary
- Bilingual transfer dictionary and rules
- Modules reusable for new language pairs

✓ Try to represent all items in input

✓ Can find distant dependencies

× Not robust if SL analysis fails

× Weak in word choice at transfer stage

Data-driven MT (SMT) is more robust but may leave bits out

Probabilities based on bilingual corpora

- 'translation model' SL word ≈ TL word
- 'language model' TL word 1 > 2 > n
- 'decoder' manages the two models

✓ Always produces some output

✓ Reads more fluently

✗ Poor on morphology, syntactic function

✗ Spurious translations or omissions

VW and EPO use rule-based MT (RMT) to make huge savings

- **VW** using *Lucy*
  De, En, Es, Fr, Ru
- 8k jobs/day (1.5k p)
- 1 machine day =
  106 translator days
- Save €50k/day
  (€19M/year)
- Plans to add Pt, Zh

- Main problem is poor ST – typos, errors
- **EPO** prime aim is {De, En, Fr} -> all EU
- Back-up for patent experts/examiners
- Domain dictionaries
- It – 30k jobs Dec'08

Google plans to make statistical MT (SMT) available everywhere

- 3 languages in 2006
  41 languages today
- Google Search, YouTube captions and RDS feeds translated on the fly
- Aim = ubiquity: text, speech, pictures on websites, mobiles …
- Large TL models (100GB) eat time
- But train eg Yiddish on 200k, bridge from De, He, Pl lexicons, morphologies
- Enrich SL analysis with En parser for SOV TLs (Hi, Ja, Ko)

SMT has been extended with knowledge-driven procedures

- For morphologically rich languages, reduce data sparseness by using lemma-based language models instead of text-form models
- Pre-process agglutinative languages (Hu, Ja, Tr …) to split complex word strings
- Include syntax in decoding operation to boost grammaticality of output
- Treat words as vectors of features: lemma, part-of-speech, morphology …

RMT has been extended with data-driven procedures

- Extract terminology from mono- or bi-lingual corpora and add to dictionaries
  - Good for filling gaps and adapting to new domains
  - Aggravates lexical choice when dictionaries are already big
- Create collocation models of SL words from corpora to enable disambiguation of words at level of paragraph rather than sentence

An RMT+SPE (statistical post-editing) 'pipeline' exists

- SPE component trained on a 'bilingual' corpus of 'raw' and post-edited text
- Significant improvements (especially word choice) in limited domains even with little training data
- Can outperform pure SMT built on little data
- SPE stage may introduce errors, eg omitting names of people or organisations

Genuine hybrids are emerging from the swamp

- METIS has tried rule-based analysis, bilingual dictionaries for transfer, and a language model based on the British National Corpus (100M)
  - Performance similar to basic SMT
  - Worse than complete RMT
  - But effort to date is relatively small
- Context-based MT tries data-driven analysis and generation, with bilingual dictionaries
  - Ungrammatical, ignores unknown words

EU is funding MT research again after a long break

EuroMatrix 2006-12

- SMT between all official EU languages
- News and web pages
- For professionals and lay users creating content in own lang.
- System learns from user corrections

FP7 ICT/4 2010-12

- Extracting terms and rules from **comparable** corpora
- Reducing size of training corpus
- Adding reliability score to translations
- MT beyond sentence

Users of MT for interaction will make more pragmatic demands

- Socialising across language barriers in virtual environments
- Need for complex user models
- Concern with affect, emotion, identity, cross-cultural etiquette …
- New metrics for evaluation from new disciplines, eg pragmatics, sociolinguistics

We will see more niche systems for specific domains and tasks

- Trained on specialised corpora
- MT embedded in information monitoring systems
- Quality measured by task performance not similarity to a human translation
- A range of evaluation metrics and standards

Thank you for listening and the interpreters for ... interpreting

Any comments or questions?