Recent Developments in Machine Translation

Elaine Uí Dhonnchadha, John Judge

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

British-Irish Council: Indigenous, Minority And Lesser Used (IML) Languages Conference

Outline

- Why MT is important/necessary
- Background (some brief history of MT)
- Approaches to MT

- Challenges for MT in general – why MT is hard
- Challenges for IML languages
- Current R&D for Irish
- Towards a joint model/system/pipeline

- Next steps

Why MT is important/necessary

What is Machine/Automatic Translation? - A process where content in one language is translated, wholly or in part, to another by some software process(s)

An enabling tool in the translation process - NOT intended to replace skilled human translators, BUT provides them with powerful tooling to increase productivity

This is important because…

1. the demand for translation has overtaken the supply of translators in EU and national government administration & product manuals and marketing information etc.
2. some translation is repetitive and routine and suitable for semi-automation

Background (some brief history of MT)

- Post war code breaking approaches
- First MT conference London in 1956
- Mid 1960's ALPAC report criticised lack of advancement. Led to funding cuts and the field stalled for some time
- 1970's Closed domain, limited scope systems - some success eg for abstracts, technical manuals
- 1980's-90's better computing power saw advances in algorithms and the ability to explore data/statistical driven approaches to MT as well as hand crafted rules.
- SYSTRAN (Babelfish) was the first free MT on the web
- Moses (2003) and other tools made large scale, reliable, useable MT a real possibility

Main approaches to MT

- Rule based approaches (RBMT)
- Word based - Bilingual dictionary and word reordering
- Linguistic analysis/generation
- Interterlingua – language independent representation
- Example-based (Multi-word based/phrase based) Statistical MT
- Statistical approaches (SMT)
- Word based / Phrase based Statistical MT
- Hybrid - Linguistically motivated SMT/Statistically enhanced RBMT

Challenges for MT in general – why MT is hard

- Word sense disambiguation
- Non standard language usage
- Named Entities
- Idiomatic expressions
- Pronouns and referential items
- Syntactic/Semantic Ambiguity (She saw the man with the telescope)
- Morphological analysis

How these challenges are being tackled?

- Real world knowledge, semantic constraints, discourse analysis
- More data – bigger corpora
- Open source software, cooperation, data sharing
- MT Competitions, Shared Tasks - encourage community to tackle problems together

*

Additional challenges for IML languages

- **Data**: lack of parallel data for statistical approaches

- **Human resources**:
  - Lack of commercial interest in IML languages means fewer people choose to work in the area
  - Lack of people with the relevant language expertise working in the area

- **Research**: Lack of linguistic research in to the IML language(s)

Open source tools and systems

Rule-based MT systems

Apertium

Spanish/Catalan and up to forty other pairs etc.

http://wiki.apertium.org/wiki/Main_Page

Demo https://www.apertium.org/index.eng.html?dir=slv-srp#translation

Matxin

Spanish/Basque etc. Also Spanish, French, Portuguese, English

http://matxin.sourceforge.net/ Demo: http://www.opentrad.com/en/

GramTrans

Danish/English , Norwegian/English, Swedish/Danish etc.

http://gramtrans.com/

OpenLogos

German ,English to French, Italian, Spanish http://logos-os.dfki.de/

Open source tools and systems

Statistical MT Tools

- **Giza++** a training tool for IBM Model 1-5 (version for gcc-4)
- **Phrasal,** a toolkit for phrase-based SMT
- **cdec,** a decoder for syntax-based SMT
- **Joshua,** a decoder for syntax-based SMT
- **Jane,** decoder for syntax-based SMT
- **Pharaoh** a decoder for phrase-based SMT
- **Rewrite** a decoder for IBM Model 4
- **BLEU scoring tool** for machine translation evaluation
- **Moses,** a complete SMT system
- **UCAM-SMT,** the Cambridge Statistical Machine Translation system

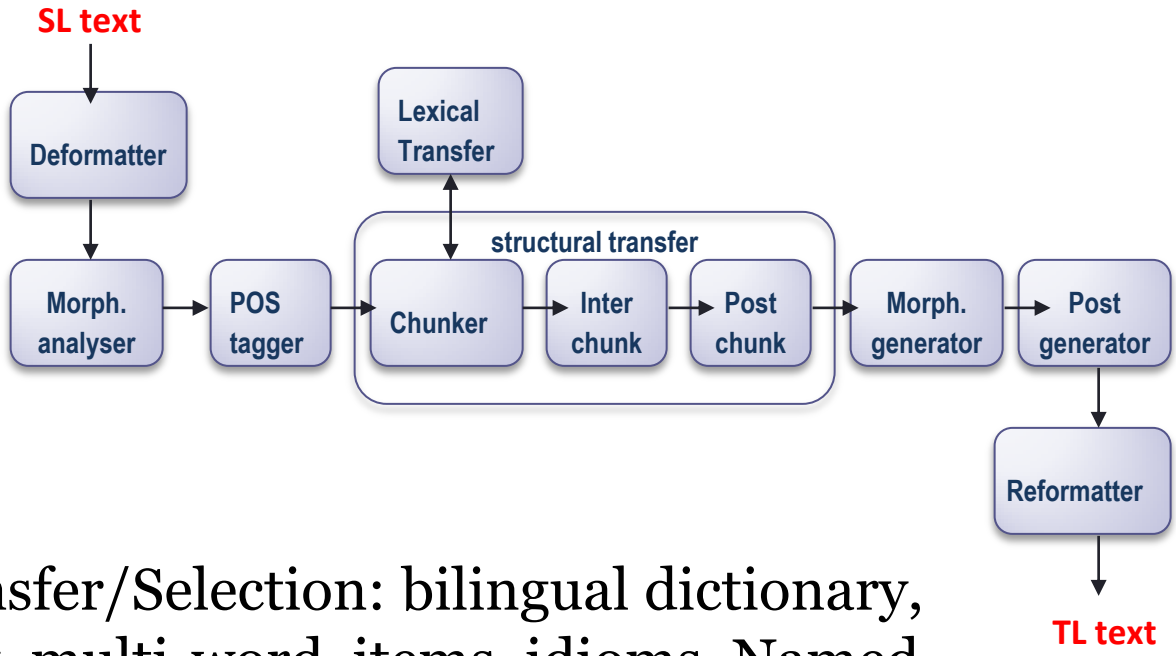
All available through http://www.statmt.org/

TCD - Developing Language Resources for Irish

- Data and Rule-based tools for the analysis and generation of Irish text
- Data: corpora – collections of written or transcribed spoken Irish, automatically annotated, gold standards, lexicons
- Text processing tools:
  Finite-state tokenisation, morphological analysis and generation, chunking
  Constraint Grammar POS tagging and partial dependency parsing

TCD – Rule Based MT

- Apertium RBMT: Irish/English, English/Irish



- Lexical Transfer/Selection: bilingual dictionary, terminology, multi-word items, idioms, Named entities (people, places etc.), corpus collocations, Wordnet-LSG,
- Structural transfer: chunking, reordering of chunks

Tapadóir - Statistical MT for Irish

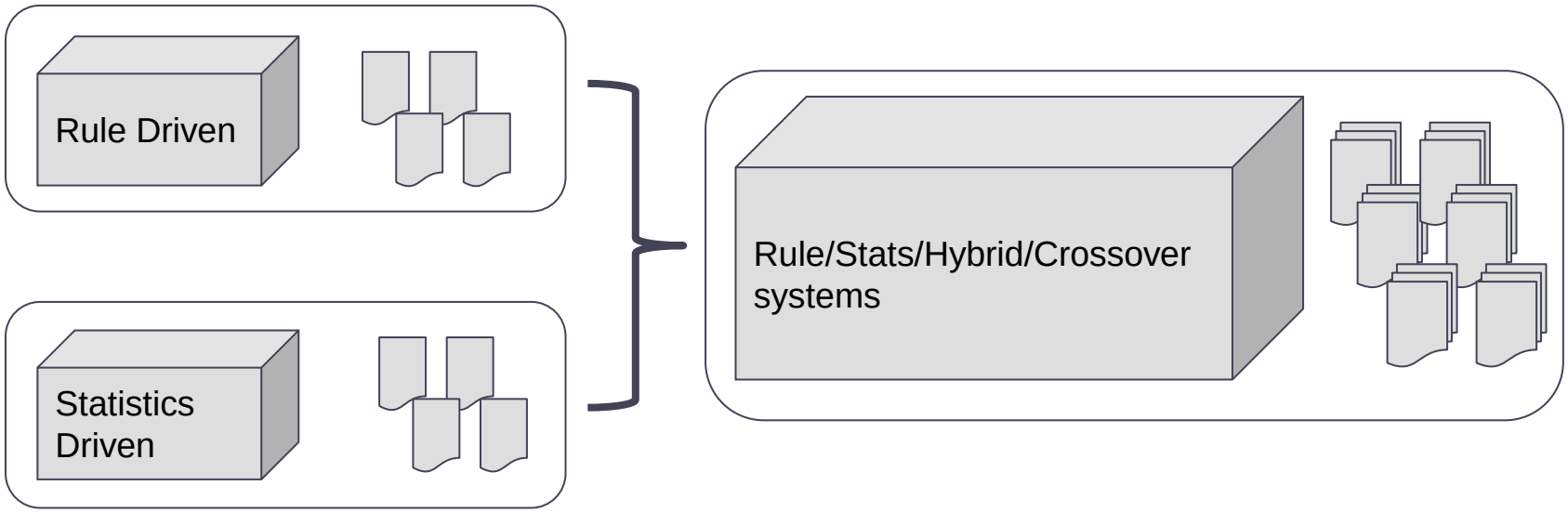
- Initial 6 month pilot to determine feasibility - moved to full scale research project
- Main Goals: Investigating use of statistical approaches for EN>GA, gathering and curating appropriate data, exploring algorithmic parameters to best fit the language(s) and use case
- Open Source Framework - Moses-based statistical engine
- Outperforms Google Translate

Tapadóir - Current Phase

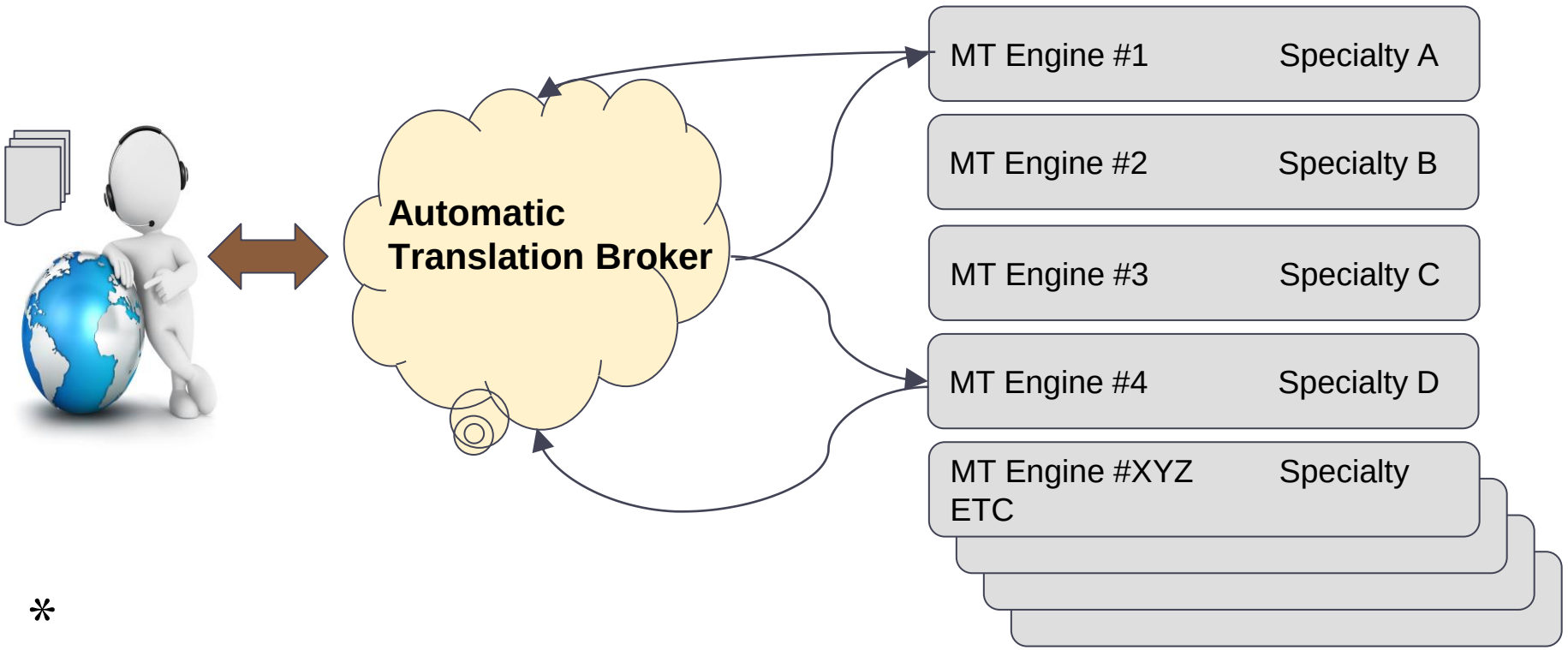
- Trials at pilot phase indicated output was useable - productivity increasing
- Early stage engine rolled out for internal use by DAHG
- Integrates seamlessly with existing translation workflow - currently translating >160k words per month
- Now: Augmenting purely statistical approach with linguistic knowledge
- Building on Elaine's work in RBMT, POS tagging and other areas to provide suitable integration points for greater integration of linguistic knowledge
- Working to enable smooth hybridisation of these approaches for optimal results

Towards a joint model/system/pipeline

- No one solution is emerging…So what next for Irish MT?
- Approaches homogenise but the applications diversify



Towards a joint model/system/pipeline



*

Next steps …

- Benefit from the work done for major languages
- SMT – language independent research
- RBMT - English is half of the translation pair … continue to develop language modules for Irish, learned rules as well as handcrafted
- Irish/IML – a good test case for research and tech.
- Share/Reuse data/software resources – dictionaries, corpora, translation memory …
- Tools for streamlining translation workflows, controlled language, house style – reduce ambiguity in MT input

Next steps …

- Build on strengths in statistical and rule-based methods and collaborate on integrated MT systems
- Look beyond our own regions - European and International initiatives
  - European Language Resource Coordination - EC action to coordinate sharing of MT resources in the Public Sector across Europe → ELRC Dublin Workshop Jan 2016
- Develop a Strategic Plan - *Plean Digiteach don Ghaeilge*

- Elaine Uí Dhonnchadha ,TCD – Dublin
- UIDHONNE@tcd.ie

- John Judge, ADAPT Centre – Dublin
- john.judge@adaptcentre.ie