# MPhil in Advanced Computer Science
# Statistical Machine Translation

| | |
|---|---|
| **Leaders:** | Dr. Stephen Clark and Dr Adrià de Gispert |
| **Timing:** | Lent |
| **Prerequisites:** | Introduction to Statistical Speech and Language Processing core module or Equivalent Background |
| **Structure:** | 10 Lectures plus 2 practicals over 6 sessions |

## AIMS

This module provides an in-depth introduction to Statistical Machine Translation, the dominant approach to providing large-scale, robust translation applicable to many language pairs (and the approach currently used by Google). Topics covered will include:

## SYLLABUS

1. Overview: (2L)
    - Translation as an economic, political, and cultural activity
    - Machine translation as a problem in natural language processing
    - Syntax and morphology in translation
    - Translation memories; example and rule-based based MT
    - Interlingua
2. Alignment: automatic translations in text (2L)
    - Parallel texts and their role in building translation systems and measuring translation quality
    - Document and sentence alignment: models and algorithms
    - Word and phrase alignment: models and algorithms
    - Techniques for automatic measurement of alignment quality
    - Webcrawling for parallel text
3. Weighted finite state transducers: algorithms for natural language processing and MT: (2L)
4. SMT Systems (4L)
    - Extraction of translation rules from parallel text
    - Phrase-based, Hiero, syntax-based MT
    - Techniques for automatic measurement of translation quality
    - Minimum error rate training

- Language models for SMT : simple back-off, MapReduce
- MT system combination
- Practical issues in SMT: true casing; source text pre-processing; handling morphology; system building procedure

All lectures will be given by Dr. Clark or Dr de Gispert.

**OBJECTIVES**

On completion of this module students should understand:

- the role of parallel text in MT
- how alignment models can be estimated from parallel text
- how alignment models capture divergent language properties such as word order
- the use of WFSTs in translation and some other basic NLP tasks
- the extraction of translation rules from parallel text
- various phrase-based translation architectures, including Hiero
- parameter optimization procedures for SMT
- the role of language models in SMT
- the evaluation of SMT systems using automatic metrics
- system combination techniques for SMT

**PRACTICAL WORK**

The will be two substantial practical exercises associated with this module.

- Practical 1: 2 sessions. Parallel text, alignment models and WFSTs
- Practical 2: 4 sessions. SMT system construction and evaluation

**ASSESSMENT**

- Written report covering the practical worth 35% of the marks.
- One final take-home exam covering all the material. Final take-home exam will contribute 65% to the final mark. Questions set and marked by Dr. Clark and Dr de Gispert.

**RECOMMENDED READING**

- SPEECH and LANGUAGE PROCESSING, Jurafsky and Martin, 2nd edition, Chapter 25 on Machine Translation
- The mathematics of statistical machine translation: Parameter estimation, PF Brown, VJ Della Pietra, SA Della Pietra, Computational linguistics, 1993
- Hierarchical phrase-based translation. David Chiang, Computational Linguistics, 33(2):201-228, 2007

Last updated: October 2009