## Module Descriptor 2016/17

School of Computer Science and Statistics.

<table>
<thead>
<tr>
<th>Module Code</th>
<th>CS4LL5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Module Name</td>
<td>Advanced Computational Linguistics: Machine Learning Techniques in Machine Translation, Speech Recognition and Topic Modelling</td>
</tr>
<tr>
<td>Module Short Title</td>
<td>Advanced Computational Linguistics</td>
</tr>
<tr>
<td>ECTS weighting</td>
<td>5</td>
</tr>
<tr>
<td>Semester/term taught</td>
<td>Michaelmas</td>
</tr>
</tbody>
</table>
| Contact Hours | Lecture hours: 22  
Lab hours: 6  
Tutorial hours: 5 |
| Module Personnel | Dr Martin Emms |

### Module Learning Aims

The aim is to give a grounding in so-called unsupervised machine learning techniques which are vital to many language-processing technologies including Machine Translation, Speech Recognition and Topic Modelling. Whilst studied in these contexts, the techniques themselves are used much more widely in data mining and machine vision for example.

- understand in general what a probabilistic model is, the distinction between so-called visible and hidden variables
- the general idea of unsupervised training as way to set model parameters concerning hidden variables from evidence only on visible variables
- understand Expectation Maximisation (EM) as a general unsupervised technique, including proofs of its convergence and property of increasing data likelihood
- understand specific instances of this in Machine Translation and Speech Recognition and further details of how seemingly infeasibly costly calculations can in fact be feasibly done
- consider the further case of models for the hidden ‘topics’ in a document collection and the further modifications to EM to solve this

### Module Content

- Probability basics on collections of variables with discrete outcomes (what word, what topic etc) in particular joint, marginal, and conditional probabilities; the chain rule; relative frequencies as maximum likelihood estimators
- Statistical Machine Translation: general (source|target) x target formulation and learning from corpus of sentence pairs; idea of ‘hidden’ alignment variables between sentence pairs; the so-called IBM alignment models; brute-force EM for learning alignment models; efficient exact algorithms avoiding the exponential cost of brute-force EM
- Speech Recognition: general Hidden Markov Model (O|S) x S formulation where O is observable speech, and S is hidden state sequence.  Brute-force EM for learning HMM parameters from corpus of observed speech; the efficient Baum-Welch algorithm avoiding the exponential cost of brute-force EM
- Topic Modelling: a technique for assisting the navigation of huge document collections by seeing them as involving hidden or latent ‘topic’ variables; how this can be used to recover hidden relationships between documents; techniques to learn parameters of these models
- in each case, alongside the explanation of the algorithms, there will be practical work, either developing instances of them, or deploying existing implementations and running them on data sets to concretely see their properties
<table>
<thead>
<tr>
<th>Module Description 2016/17</th>
</tr>
</thead>
<tbody>
<tr>
<td>School of Computer Science and Statistics.</td>
</tr>
</tbody>
</table>

### Recommended Reading List

I will be providing notes, sometimes directing attentions to particular chapters from the following books, as well as possible online sources:

- Kevin Murphy's book 'Machine Learning: A Probabilistic Perspective'
- Russel and Norvig's book 'Artificial Intelligence: A Modern Approach'
- Jurafsky and Martin's book 'Speech and Language Processing'
- Phillip Koehn's book 'Statistical Machine Translation'
- Manning and Schutze's book 'Foundations of Statistical Natural Language Processing'

### Module Pre-Requisite

No pre-requisite: to implement and experiment with tools will need to be able to program in C++

### Module Co-Requisite


### Assessment Details

- Examination: 70%
- Course Work: 30%

Marks simply combined: not required to pass each component separately.
Exam takes place in January.

### Module Approval Date

Approved By

### Academic Start Year

2014-2015

### Academic Year of Data