Module Code	CSU44062
Module Name	Advanced Computational Linguistics: Machine Learning Techniques in Machine Translation, Speech Recognition and Topic Modelling
ECTS Weighting [1]	5 ECTS
Semester taught	Semester 1
Module Coordinator/s	Dr Martin Emms
Module Learning Outcomes	When students have successfully completed this module they should be able to: • LO1 understand in general what a ***probabilistic model*** is, the distinction between so-called ***visible and hidden*** variables, and the distinctive nature of models where each datum is ***a sequence of varying length***, rather then a ***fixed-size set of features*** • LO2 understand the general idea of ***unsupervised training*** as way to set model parameters concerning hidden variables from evidence only on visible variables • LO3 understand ***Expectation Maximisation (EM)*** as a general unsupervised technique, including proofs of its convergence and property of increasing data likelihood • LO4 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 • LO5 consider the further case of models for the ***hidden 'topics' in a document collection*** and the further modifications to EM to solve this • LO6 the aim is to give a grounding in so-called unsupervised machine learning techniques, vital to many language-processing technologies. Whilst most time will be directed to these contexts, the techniques themselves are used much more widely in data mining and machine vision for example, and students will gain some insight into this also.

[1] TEP Glossary

Module Content	1. Probability basics on collections of variables with discrete outcomes (what word, what topic etc) in particular joint, marginal, and conditional probabilities; the chain rule; 2. Idea of machine learning as ***parameter estimation to maximise likelihood*** of training data. Case of ***supervised*** parameter estimations, with all model variables ***visible*** in data; illustrations/proofs that intuitive relative frequency approaches are maximum likelihood estimators. Case of ***unsupervise***d parameter estimation, with some model variables ***hidden*** in data. Details of how ***EM algorithm*** achieves seemingly impossible feat of finding likelihood maximising estimate in this case, with hidden variables ***summed over***. 3. ***Statistical Machine Transation***: 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 4. Generalisation to so-called '***Phrase-based SMT***'. Dealing with the algorithmic challenges of 'decoding' ie. finding a best solution 5. ***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; exploration of how the efficient Baum-Welch algorithm achieves seemingly impossible feat of avoiding the exponential cost of brute-force EM 6. ***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 7. 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
Teaching and Learning Methods	There is a mixture of lectures, tutorials and lab sessions. Most frequently there will be a 3 lectures per week, but there will be occassions where 1 or more of the time-tabled lecture sessions will actually be a lab-session or a tutorial. This may happen in anticipation of the setting of a course work assignment. In some cases there will be alternate forms of an assignment, one form more mathematical requiring 'pen-and-pencil' calculations, the other more implementational requiring the implementation of some (part of) some algorithm. There will be many further exercises in online materials, all of which students will be encouraged to attempt; To all of the exercises suggested answers will be provided some time after the exercise has been first made available

Assessment Details
[2]

Assessment Component	Brief Description	Learning Outcomes Addressed	% of total	Week set	Week due
Examination	2 hour examination examination	LO1, LO2, LO3, LO4, LO5, LO6	70%	n/a	n/a
Course work		LO1, LO2, LO3, LO4, LO5, LO6	30%	N/a	N/a

Reassessment Details

Examination (2 hours, 100%)

Contact Hours and Indicative Student Workload

Contact Hours (scheduled hours per student over full module), broken down by:	**33 hours**
lecture	22 hours
laboratory	6 hours
tutorial or seminar	5 hours
other	0 hours
Independent study (outside scheduled contact hours), broken down by:	**36 hours**
preparation for classes and review of material (including preparation for examination, if applicable)	18 hours
completion of assessments (including examination, if applicable)	18 hours
Total Hours	**69 hours**

Recommended Reading List

Online notes will be providing notes. Sometimes these will directing attention to particular chapters from the following books, as well as possible online sources

1. Jurafsky and Martin's book 'Speech and Language Processing'
2. Russel and Norvig's book 'Artificial Intelligence: A Modern Approach'
3. Phillip Koehn's book 'Statistical Machine Translation' (associated site: www.statmt.org/book)
4. Kevin Murphy's book 'Machine Learning: A Probabilistic Perspective'
5. Witten and Frank's book 'Data Mining Practical Machine Learning Tools and Techniques'
6. online Michael Collins *about EM* http://www.cs.columbia.edu/~mcollins/6864/slides/em1.4up.pdf
7. online Do & Batzoglou *about EM* http://ai.stanford.edu/~chuongdo/papers/em_tutorial.pdf
8. online Steyvers and Griffiths(2006) *about Probabilistic topic models*. http://cocosci.berkeley.edu/tom/papers/SteyversGriffiths.pdf

[2] TEP Guidelines on Workload and Assessment

Module Pre-requisites	**Prerequisite modules:** none **Other/alternative non-module prerequisites:** the module is self-contained. To be able to do some variants of assignments, ability to program is required. As noted above, there will be an alternative non-programming more maths-based version of any assignment involving programming.
Module Co-requisites	
Module Website	www.scss.tcd.ie/Martin.Emms/4CSLL5/
Last Update	24/07/2019 by Martin Emms