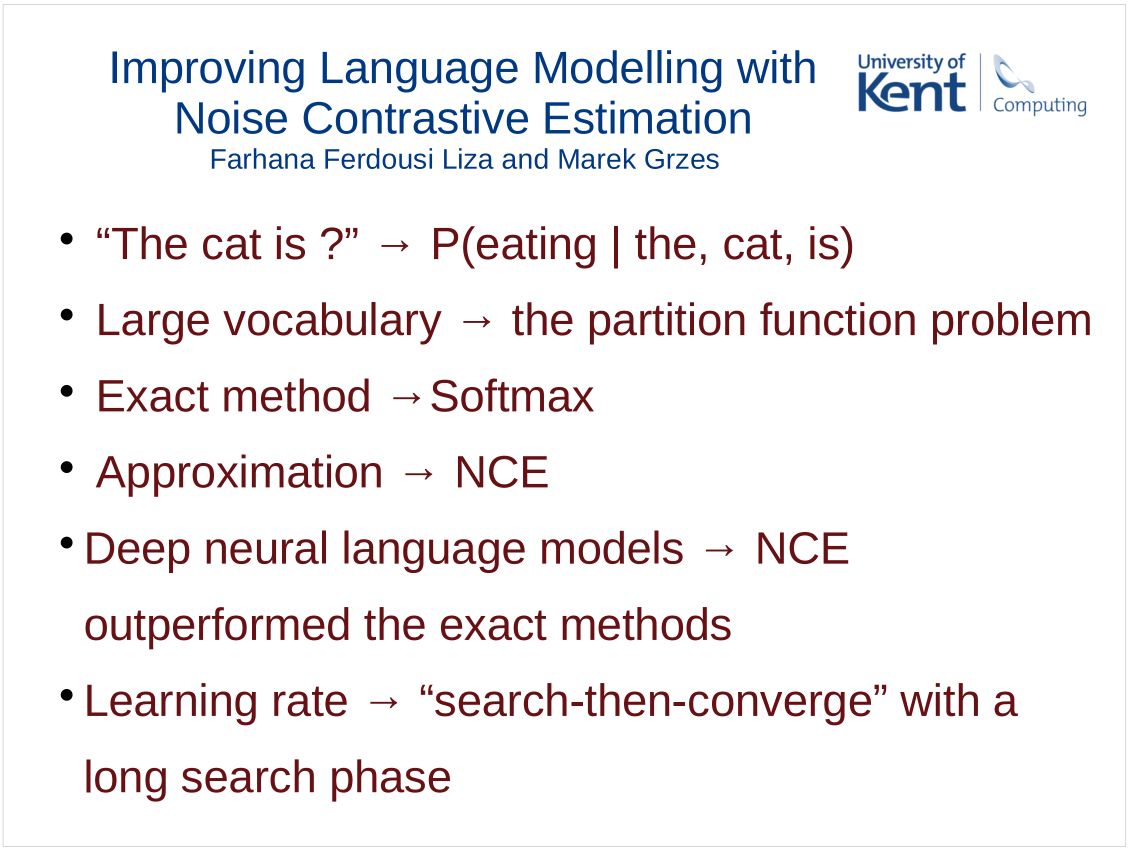

Improving Language Modelling with Noise Contrastive Estimation

Farhana Ferdousi Liza and Marek Grzes

- “The cat is ?” → P(eating | the, cat, is)
- Large vocabulary → the partition function problem
- Exact method → Softmax
- Approximation → NCE
- Deep neural language models → NCE outperformed the exact methods
- Learning rate → “search-then-converge” with a long search phase

- In language modelling, we are interested in the probability of an upcoming word.
- The probability distribution is usually over many words; therefore, the normalisation constant is a bottleneck.
- An exact method to normalise the probability distribution over words is to use softmax
- NCE is one of the approximate methods
- The objective function is highly non-convex in deep learning; different components of the algorithms can improve convergence to a better local optima
- Potentially because of the reason mentioned above, in our paper, we showed, for the first time, that NCE can beat softmax on domains on which softmax is known to perform well, and we achieved the best results in the class of algorithms with standard dropout and standard LSTM
- One of the key ideas that allowed us to achieve these results was the “search-then-converge” learning rate schedule and the fact that the search phase has to be sufficiently long in NCE.