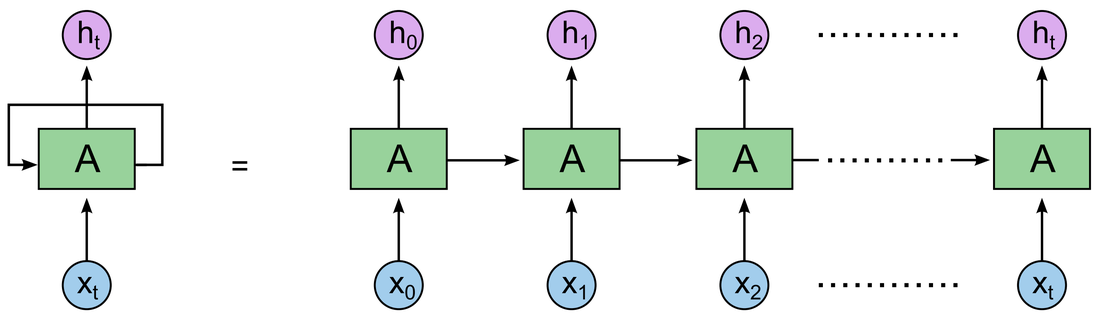
UNIVERSITY OF COPENHAGEN, DENMARK
AALBORG UNIVERSITY COPENHAGEN, DENMARK

# Deep Learning Relevance: Creating Relevant Information (As Opposed to Retrieving It)

Christina Lioma, Department of Computer Science, University of Copenhagen, Denmark
Birger Larsen, Department of Communication, Aalborg University Copenhagen, Denmark
Casper Petersen, Department of Computer Science, University of Copenhagen, Denmark
Jakob Grue Simonsen, Department of Computer Science, University of Copenhagen, Denmark

**What if Information Retrieval (IR) systems did not just retrieve relevant information that is stored in their indices, but could also *understand* it and *synthesise* it into a *single* document?**

**First step in this direction: use deep learning to emulate "learning the semantics" of all indexed relevant information for a query and compose a new document conveying those semantics.**




**1. Use Long Short Term Memory (LSTM) Recurrent Neural Networks (RNN) to learn relevant information on a character level (3 layers, 512 neurons).**

Distributional Semantics Hypothesis:
*"You shall know a word by the company it keeps"*
J. R. Firth

**2. Learn only the *distributional semantics* of the indexed relevant information to reduce learning noise.**

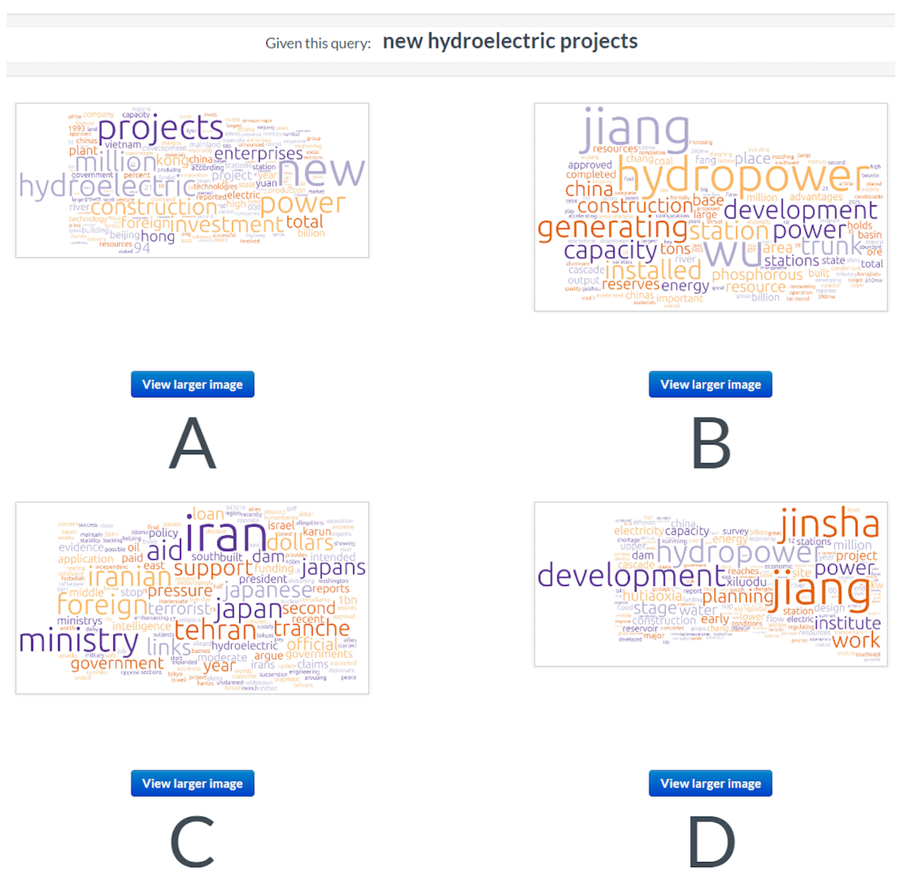
METHODOLOGY

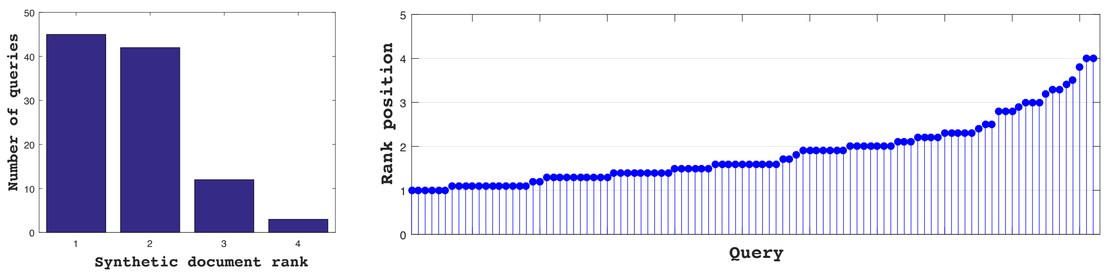
**Input**: a query and its relevant documents

1. Extract context windows (CW) of ±30 terms around each query term from all relevant documents
2. Train a LSTM RNN on the characters from the extracted CWs
3. Sample characters from the trained LSTM RNN

**Output**: A synthetic document d

**3. Experiments with TREC Disks 4&5, queries 301-450 (only 101 had enough training data).**
**For each query, ask crowdsourced users to rank wordclouds of three randomly chosen known relevant documents and the RNN-created document.**




**4. Finding: the wordclouds of documents synthesised by deep learning were assessed by users as overall most relevant (average rank position 1.8).**

Work partially supported by the Danish Industry Foundation and the IDAS (Industrial Data Analysis Service) project.