Key Point Extraction

Automating Highlight Generation

December 2019 – Lancaster University

Daniel Kershaw

Outline

- Product ideation
- Summarization
- Data
- RNN & LSTMS
- Model
- Evaluation
- Sentence Simplification
- Production
- SME Evaluation

Research Lead by Product Needs

Outline

Show full outline

Figures (3)

Tables (3)

Table 1
Table 2
Table 3

ELSEVIER

Knowledge-Based Systems
Volume 21, Issue 3, April 2008, Pages 192-199

A flexible framework to experiment with ontology learning techniques

Ricardo Gacitua, Pete Sawyer, Paul Rayson

Show more

https://doi.org/10.1016/j.knosys.2007.11.009

Abstract

Ontology learning refers to extracting conceptual knowledge from several sources and building an ontology from scratch, enriching, or adapting an existing ontology. It uses methods from a diverse spectrum of fields such as natural language processing, artificial intelligence and machine learning. However, a crucial challenging issue is to quantitatively evaluate the usefulness and accuracy of both techniques and combinations of techniques, when applied to ontology learning. It is an interesting problem because there are no published comparative studies.

We are developing a flexible framework for ontology learning from text which provides a cyclical process that involves the successive application of various NLP techniques and learning algorithms for concept extraction and ontology modelling. The framework provides support to evaluate the usefulness and accuracy of different techniques and possible combinations of techniques into specific processes, to deal with the above challenge. We show our framework's efficacy as a workbench for testing and evaluating concept identification. Our initial experiment supports our assumption about the usefulness of our approach.

Part of special issue:
AI 2007
Edited by Max Bramer

Download full issue

Other articles from this issue

An evolutionary algorithm-based approach to r...
April 2008, pp.
Download PDF View details

Merging intelligent agency and the Semantic Web
April 2008, pp.
Download PDF View details

Modelling affective-based music compositional ...
April 2008, pp.
Download PDF View details

View more articles

Recommended articles

Citing articles (48)

Article Metrics

Citations

Citation Indexes: 48

Captures

Available online at www.sciencedirect.com
ScienceDirect
Knowledge-Based Systems 21 (2008) 192–199
Knowledge-Based SYSTEMS
www.elsevier.com/locate/knosys

A flexible framework to experiment with ontology learning techniques

Ricardo Gacitua *, Pete Sawyer, Paul Rayson

Lancaster University, Computing Department, Infolab21, South Drive, LA1 4WA Lancaster, UK

Available online 23 November 2007

Abstract

Ontology learning refers to extracting conceptual knowledge from several sources and building an ontology from scratch, enriching, or adapting an existing ontology. It uses methods from a diverse spectrum of fields such as natural language processing, artificial intelligence and machine learning. However, a crucial challenging issue is to quantitatively evaluate the usefulness and accuracy of both techniques and combinations of techniques, when applied to ontology learning. It is an interesting problem because there are no published comparative studies.

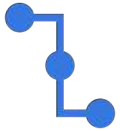
We are developing a flexible framework for ontology learning from text which provides a cyclical process that involves the successive application of various NLP techniques and learning algorithms for concept extraction and ontology modelling. The framework provides support to evaluate the usefulness and accuracy of different techniques and possible combinations of techniques into specific processes, to deal with the above challenge. We show our framework's efficacy as a workbench for testing and evaluating concept identification. Our initial experiment supports our assumption about the usefulness of our approach.

Keywords: Semantic Web; Ontologies; Ontology learning; NLP methods; Machine learning methods

Data Science Path

Extract

Extract key points from a document e.g. main findings, methods and results



Connect

Connect these to core locations within the document

Relate

Find relations between extracted sentences across documents - OpenIE

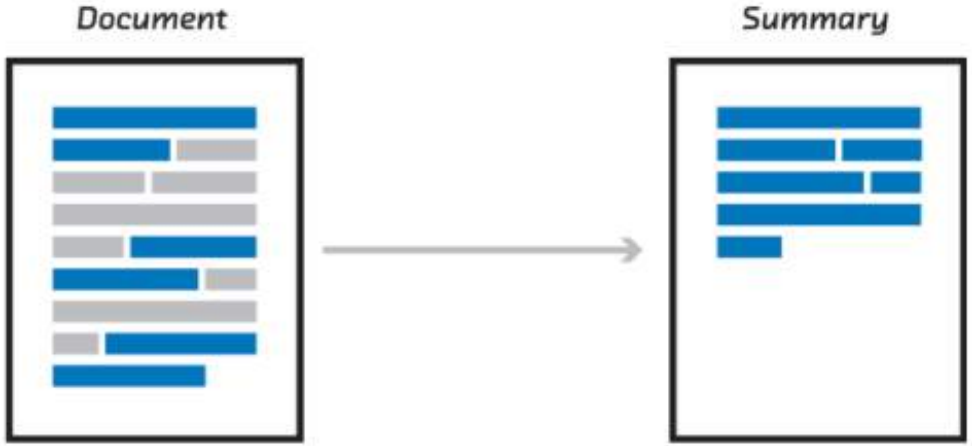
Summarization for Key point Extraction

Text summarization is the technique for generating a concise and precise summary of voluminous texts while focusing on the sections that convey useful information, and without losing the overall meaning.

1. Summaries reduce reading time.
2. Automatic summarization improves the effectiveness of indexing.
3. Automatic summarization algorithms are less biased than human summarizers.
4. Personalized summaries are useful in question-answering systems as they provide personalized information.

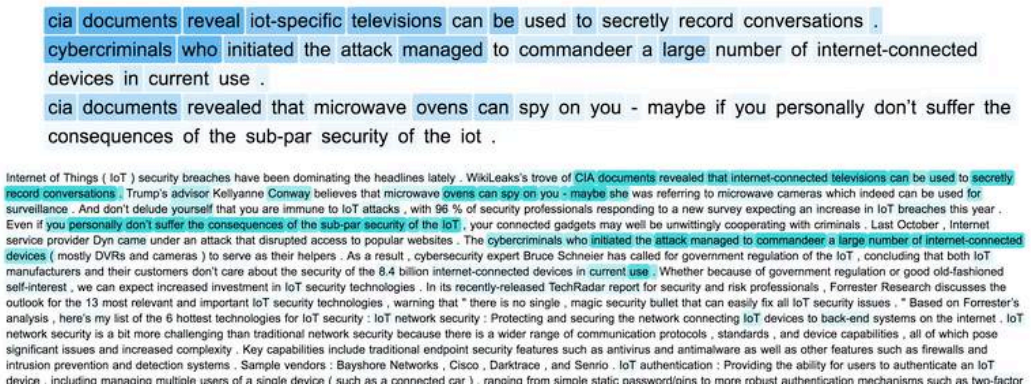
Extractive Summarization

- Select Spans of text which are summary ”like”
- No rewriting of text
- Use author sentences
- Examples: key phrase extraction, key clauses, sentences or paragraphs



Abstractive Summarization

- Involves paraphrasing of source document
- Condense text down more strongly than extractive
- Seq2seq models



Can we use extractive summarization to find the key finding/points within a document

Data

Available Data

Full Text

Australia, and several other industrialized nations, require an extensive science, technology, engineering, and mathematics (STEM) workforce for economic prosperity, productivity, and global competitiveness. However, the demand for people in STEM outweighs the supply of STEM-trained individuals. One reason for this supply-demand issue is a decline in the proportion of students choosing STEM-related pathways (Ainley, Kos, & Nicholas, 2008). In response to this concern, bourgeoning research has been devoted to identifying predictors of STEM educational and career choices (Shoffner & Dockery, 2015). Among the determinants examined is vocational interests (Bartlett, Perera, & McIlveen, 2016), which is unsurprising, given not only theory positing a central role of interests in choice behaviors (Lent, Brown, & Hackett, 1994) but also extant evidence demonstrating that interests predict choices (Gasser et al., 2007, Larson et al., 2010, Päßler and Hell, 2012). However, existing research, with few exceptions (Leuty et al., 2016, McLarnon et al., 2015), is limited to investigating the unique and additive relations of interests with choices from a variable-centered perspective. This approach assumes that individuals in a sample are from the same population and share the same set of parameters, disregarding the potential existence of multiple latent subpopulations that may show distinct configurations of interests. The near-exclusive focus on unique relations is problematic given work showing that individuals may simultaneously endorse multiple interests (McLarnon et al., 2015, Strahan and Severinghaus, 1992, Tay et al., 2011). From a social cognitive perspective on the career choice process, such interest combinations may be more important for people's educational and vocational choices than interests in isolation and may be a truer representation of individuals' interest profiles, which themselves emerge, in part, from people's dispositional characteristics. However, only little research has been conducted to determine how interests can be combined, and even less is known about how these combinations predict individuals' choices and are predicted by theoretically-meaningful antecedents in the career choice process, such as personality dispositions.

Available Data

Title

A social influence model of consumer participation in network- and small-group-based virtual communities

Abstract

We investigate two key group-level determinants of virtual community participation —group norms and social identity—and consider their motivational antecedents and mediators.

We also introduce a marketing-relevant typology to conceptualize virtual communities, based on the distinction between *network-based* and *small-group-based* virtual communities. Our survey-based study, which was conducted across a broad range of virtual communities, supports the proposed model and finds further that virtual community type moderates consumers' reasons for participating, as well as the strengths of their impact on group norms and social identity. We conclude with a consideration of managerial and research implications of the findings.

Available Data

Keywords

Vocational interests; Interest profiles; STEM career choices; Academic and career choices; Latent profile analysis; Profile invariance; Profile similarity

Article Metrics

Citations

Citation Indexes:	6

Captures

Exports-Saves:	18
Readers:	43

Social Media

Tweets:	13

References

Ainley et al., 2008 J. Ainley, J. Kos, M. Nicholas
Participation in science, mathematics, and technology in Australian education
ACER Research Monograph (No. 63)
(2008)
http://research.acer.edu.au/acer_monographs/4
Google Scholar

Ainley et al., 1990 J. Ainley, W. Jones, K.K. Navaratnam
Subject choice in senior secondary school
Australian Publishing Service, Canberra, ACT (1990)
Google Scholar

Armstrong and Vogel, 2009 P.I. Armstrong, D.L. Vogel
Interpreting the interest–efficacy association from a RIASEC perspective
Journal of Counseling Psychology, 56 (3) (2009), pp. 392-407, 10.1037/a0016407
CrossRef View Record in Scopus Google Scholar

Armstrong et al., 2008 P.I. Armstrong, W. Allison, J. Rounds
Development and initial validation of brief public domain RIASEC marker scales
Journal of Vocational Behavior, 73 (2) (2008), pp. 287-299
https://doi.org/10.1016/j.jvb.2008.06.003
Article Download PDF View Record in Scopus Google Scholar

Asparouhov and Muthén, 2014 T. Asparouhov, B. Muthén
Auxiliary variables in mixture modeling: Three-step approaches using M plus
Structural Equation Modeling: A Multidisciplinary Journal, 21 (3) (2014), pp. 329-341
https://doi.org/10.1080/10705511.2014.915181
CrossRef View Record in Scopus Google Scholar

Available Data

Highlights

- Latent profiles of vocational interests were identified.
- The profiles replicated across subsamples.
- Big-Five personality dimensions differentiated the profiles.
- Profile membership was associated with the probability of STEM major choice.

Focusing of text

Paper

Australia, and several other industrialized nations, require an extensive science, technology, engineering, and mathematics (STEM) workforce for economic prosperity, productivity, and global competitiveness. However, the demand for people in STEM outweighs the supply of STEM-trained individuals. One reason for this supply-demand issue is a decline in the proportion of students choosing STEM-related pathways (Ainley, Kos, & Nicholas, 2008). In response to this concern, bourgeoning research has been devoted to identifying predictors of STEM educational and career choices (Shoffner & Dockery, 2015). Among the determinants examined is vocational interests (Bartlett, Perera, & McIlveen, 2016), which is unsurprising, given not only theory positing a central role of interests in choice behaviors (Lent, Brown, & Hackett, 1994) but also extant evidence demonstrating that interests predict choices (Gasser et al., 2007, Larson et al., 2010, Päßler and Hell, 2012). However, existing research, with few exceptions (Leuty et al., 2016, McLarnon et al., 2015), is limited to investigating the unique and additive relations of interests with choices from a variable-centered perspective. This approach assumes that individuals in a sample are from the same population and share the same set of parameters, disregarding the potential existence of multiple latent subpopulations that may show distinct configurations of interests. The near-exclusive focus on unique relations is problematic given work showing that individuals may simultaneously endorse multiple interests (McLarnon et al., 2015, Strahan and Severinghaus, 1992, Tay et al., 2011). From a social cognitive perspective on the career choice process, such interest combinations may be more important for people's educational and vocational choices than interests in isolation and may be a truer representation of individuals' interest profiles, which themselves emerge, in part, from people's dispositional characteristics. However, only little research has been conducted to determine how interests can be combined, and even less is known about how these combinations predict individuals' choices and are predicted by theoretically-meaningful antecedents in the career choice process, such as personality dispositions.

Abstract

Abstract

We investigate two key group-level determinants of virtual community participation —group norms and social identity—and consider their motivational antecedents and mediators.

We also introduce a marketing-relevant typology to conceptualize virtual communities, based on the distinction between *network-based* and *small-group-based* virtual communities. Our survey-based study, which was conducted across a broad range of virtual communities, supports the proposed model and finds further that virtual community type moderates consumers' reasons for participating, as well as the strengths of their impact on group norms and social identity. We conclude with a consideration of managerial and research implications of the findings.

Author Highlights

Highlights

- Latent profiles of vocational interests were identified.
- The profiles replicated across subsamples.
- Big-Five personality dimensions differentiated the profiles.
- Profile membership was associated with the probability of STEM major choice.

Can we predict which sentences are most like highlights?

ELSEVIER

Sampling

Positive: 10 random samples from the top 10% of most similar sentences to highlights using rouge-l-f

Negative: 10 random samples from the bottom 10% of most similar sentences to highlights using rouge-l-f

Rouge

$$ROUGE - N = \frac{\sum_{S \in S_H} \sum_{G_n \in S} C_m(g_n)}{\sum_{S \in S_H} \sum_{g_n \in S} C(g_n)}$$

S_H is the set of manual summaries (target)
S is an individual summery
g_n is an N-gram
$C(g_n)$ is the number of co-ocurrances of g_n in the manual and automatic summary

Rouge

Rouge-*recall* - This means that all the words in the reference summary has been captured by the system summary,

Rouge-*precision* - what you are essentially measuring is, how much of the system summary was in fact relevant or needed?

Bioactivity-guided cut countercurrent chromatography for isolation of lysine-specific demethylase 1 inhibitors from *Scutellaria baicalensis* Georgi

Chao Han [a, 1], Shanshan Wang [a, 1], Zhongrui Li [b], Chen Chen [a], Jiqin Hou [a], Dingqiao Xu [a], Ruizhi Wang [a], Yaolan Lin [a], Jianguang Luo [a], Lingyi Kong [a]

Show more

https://doi.org/10.1016/j.aca.2018.01.014

Highlights

- A bioactivity-guided cut CCC strategy was designed.
- Gradient-elution CCC with real-time detection of LSD1 inhibition was developed.
- Online-storage recycling CCC was designed to obtain inhibitors in one apparatus.
- Six natural LSD1 inhibitors were isolated from *S. baicalensis* Georgi.

Example Samples

1. In order to enhance the efficiency of the discovery of natural active constituents from plants, a bioactivity-guided cut CCC separation strategy was developed and used here to isolate LSD1 inhibitors from S. baicalensis Georgi.
2. Here, fractions A (retention time: 0–200 min), B (245–280 min) and C (317–622 min) were discard because their LSD1 inhibition ratio was <50%, whereas fractions 1 (200–245 min) and 2 (280–317 min) were retained because their LSD1 inhibition ratio >50% (Fig. 2(a) and (b)), and these two fractions were stored in coil I by switching on the six-port valve I (Fig. 1(b)).
3. Gradient-elution CCC coupled with real-time detection of inhibitory activity in the collected fractions was first established to accurately locate active fractions.
4. 'However, the bioactivity-guided cut HSCCC separation method that we have developed can efficiently separate all the fractions and thus enable the purification of constituent compounds in one step by using a single CCC apparatus.
5. The LSD1 inhibitory activities of the target-isolated flavones 1–6 were evaluated to obtain their IC50 values (Table 2, Fig. S19–S24).
6. Thus, the natural LSD1 inhibitors 1-6 were successfully isolated using the bioactivity-guided cut CCC separation mode in a single step from the crude extract of S. baicalensis Georgi (Fig. 1 and 2)

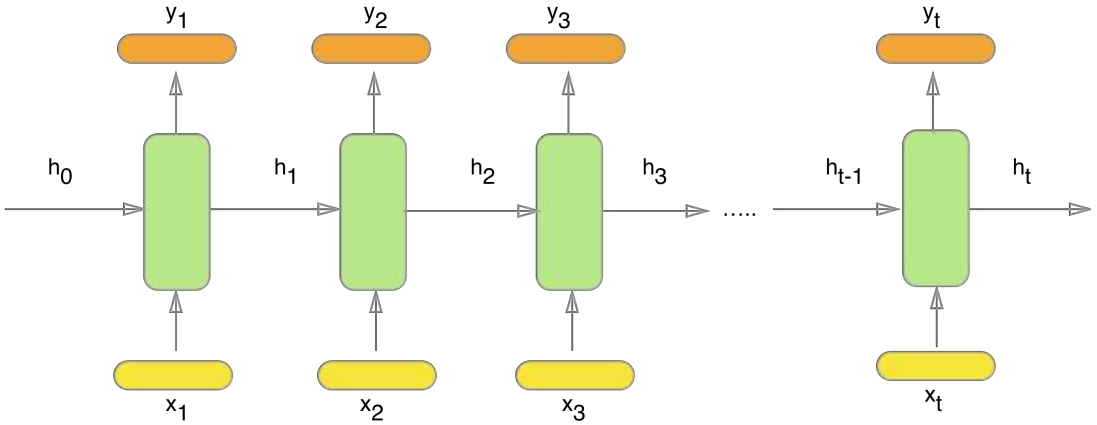
Modeling

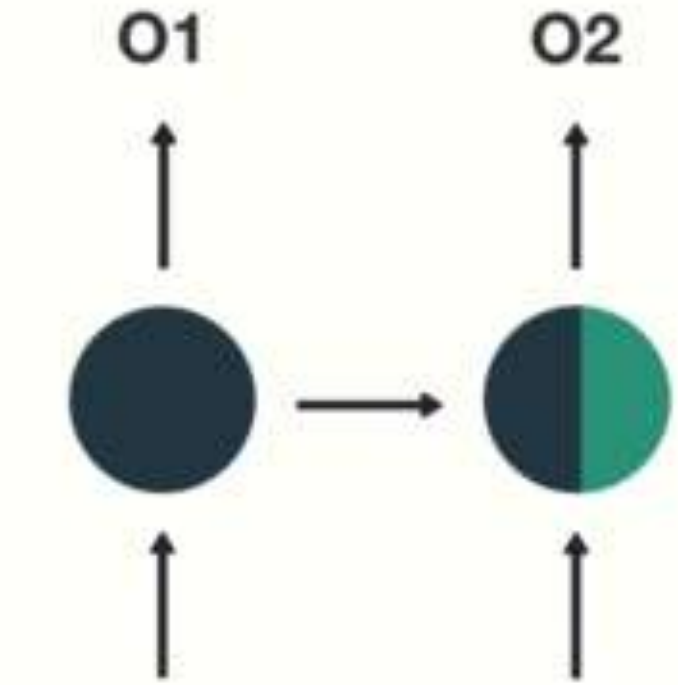
Model

- Given a sequence of words can we classify the whole sequence as a highlight
- The model needs to take the sequence into account (RNN/LSTM)
- Wanted to test out Deep Learning

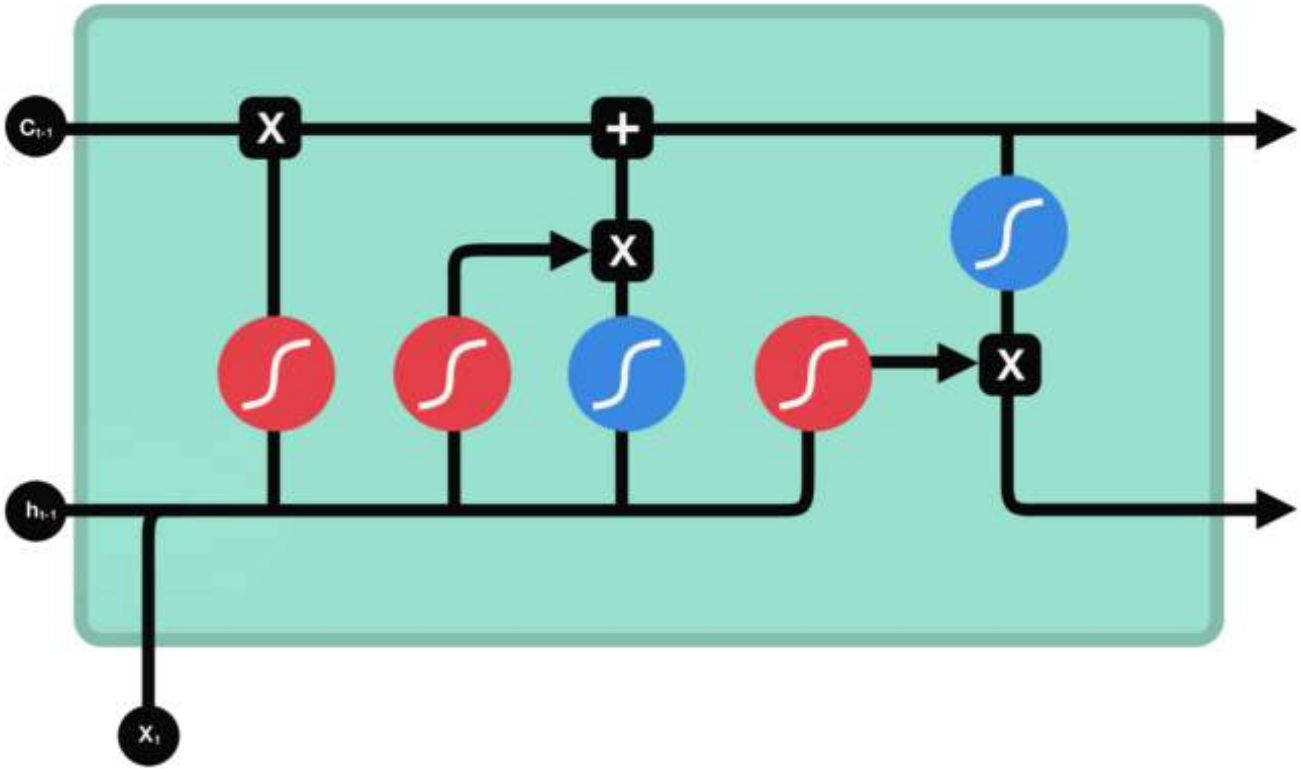
RNN

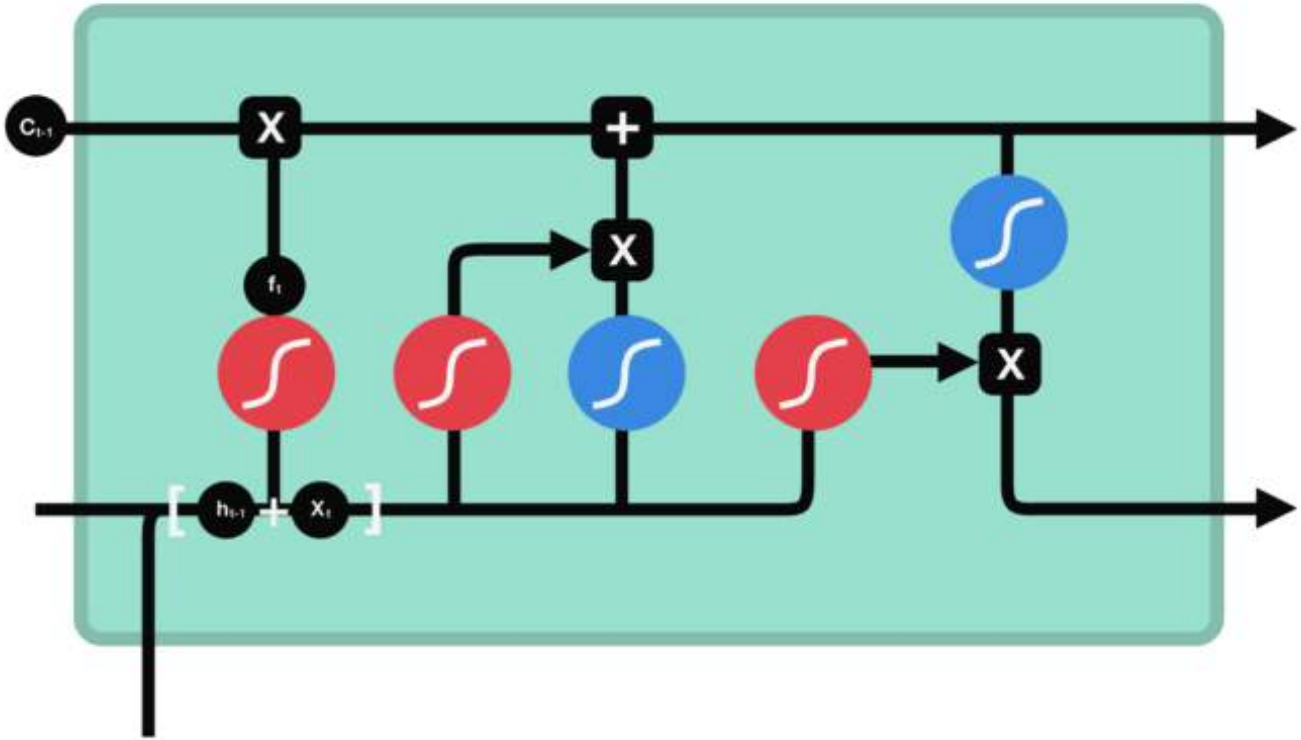
RNN networks have difficulty memorizing words from far away in the sequence





What time is it ?


c_{t-1}
h_{t-1}
x_t
X
+
X
X
c_{t-1} previous cell state
f_t forget gate output

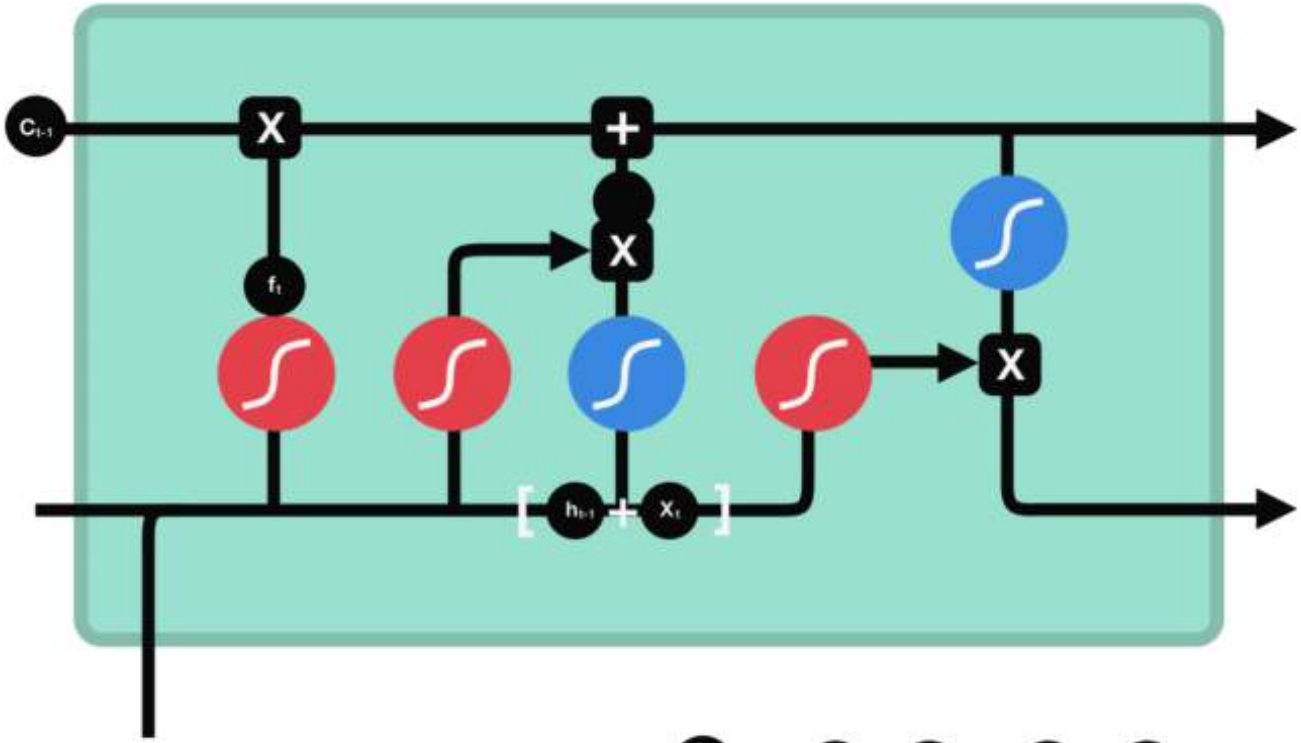


C_{t-1} previous cell state

f_t forget gate output

i_t input gate output

$\tilde{C}_t$ candidate



C_{t-1} previous cell state

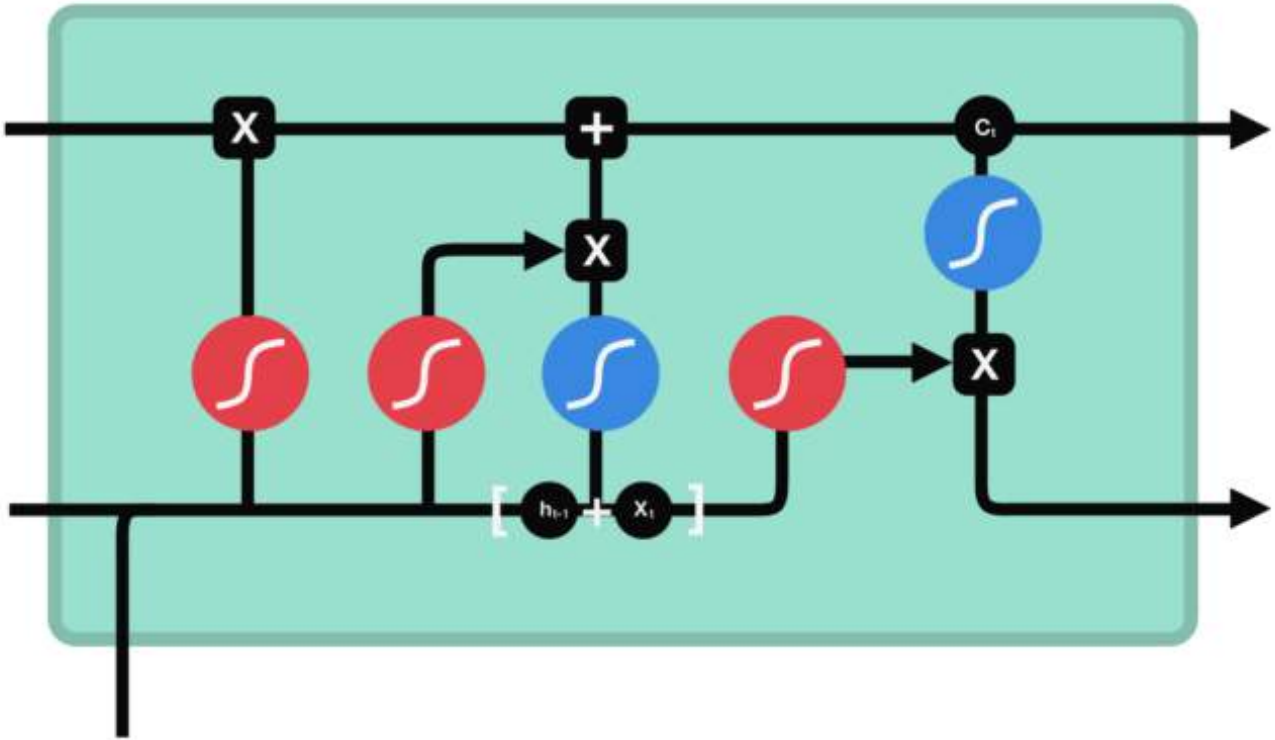
f_t forget gate output

i_t input gate output

$\tilde{C}_t$ candidate

C_t new cell state

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$



c_{t-1} previous cell state

f_t forget gate output

i_t input gate output

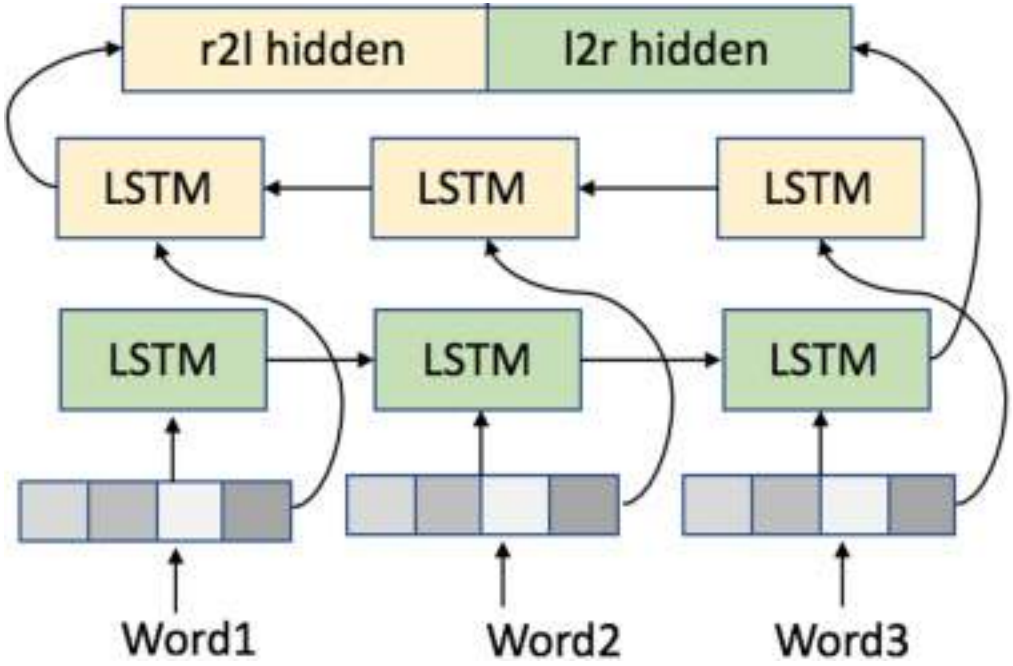
$\tilde{c}_t$ candidate

c_t new cell state

o_t output gate output

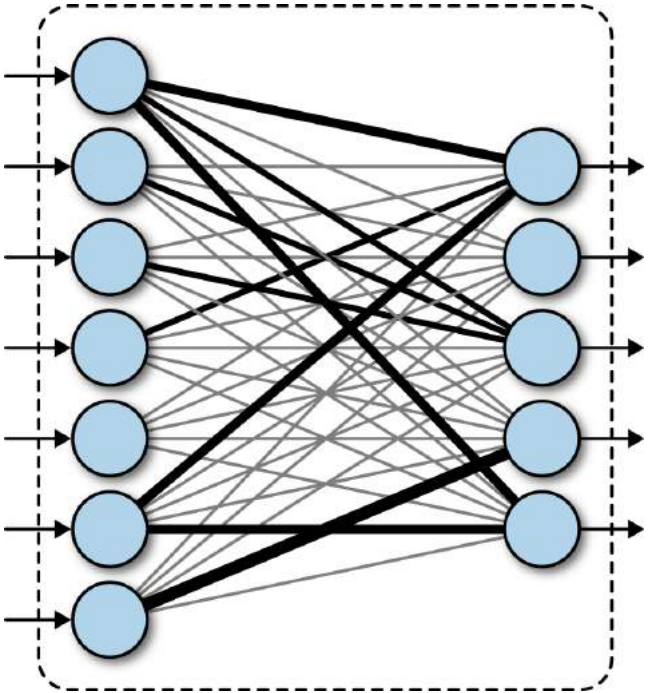
h_t hidden state

Bi-directional LSTM



Fully Contented Layer

Fully connected layers connect every neuron in one **layer** to every neuron in another **layer**. It is in principle the same as the traditional multi-**layer** perceptron neural network (MLP).

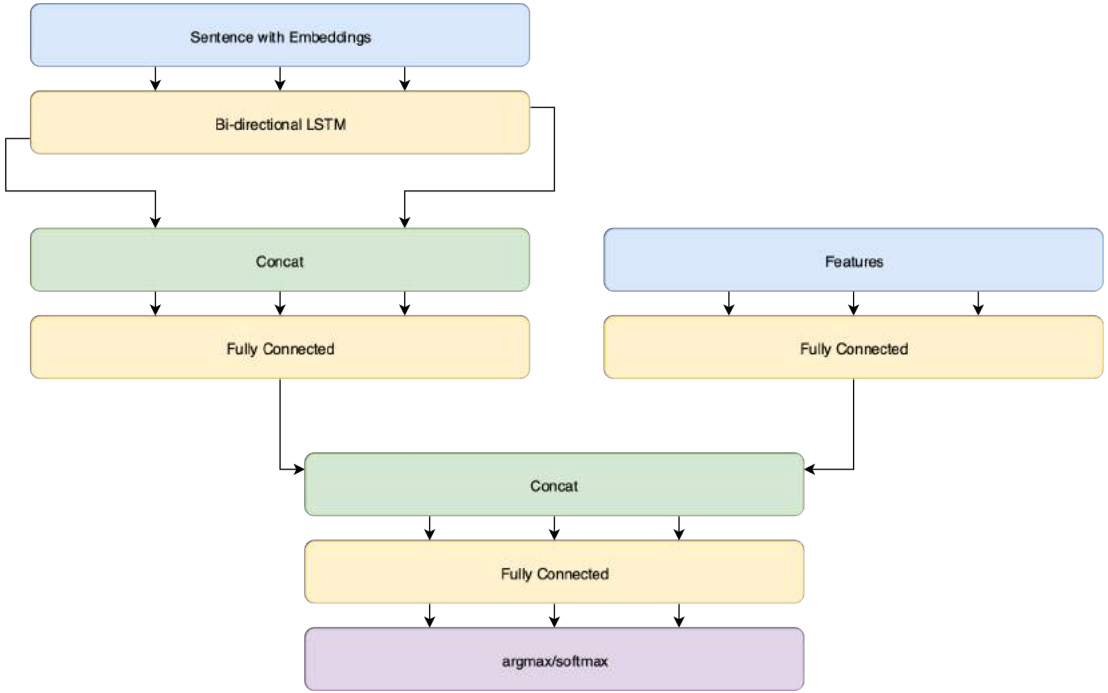


Additional Features

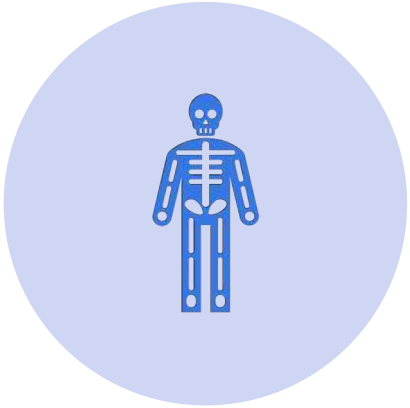
- Sentence overlap with title (number)
- Abstract embedding (sum of word embeddings)
- Journal Classifications (one hot encoding)
- Number of numbers in sentence (number)
- And some others

- All concatenated into one large feature vector

Final Model


Sentence with Embeddings
Bi-directional LSTM
Concat
Fully Connected
Features
Fully Connected
Concat
Fully Connected
argmax/softmax

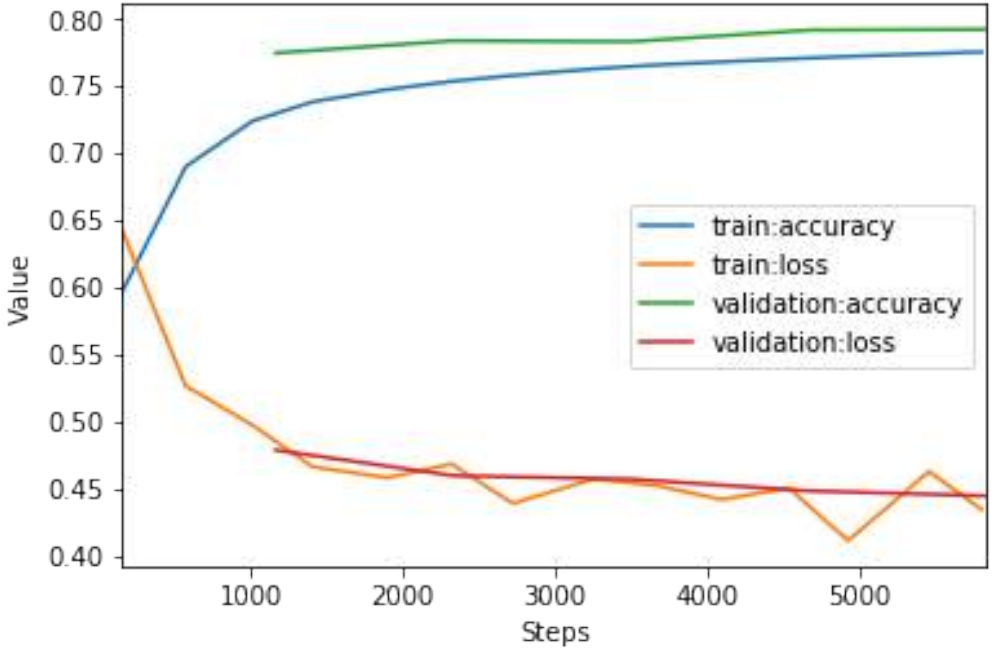
Objective Measure

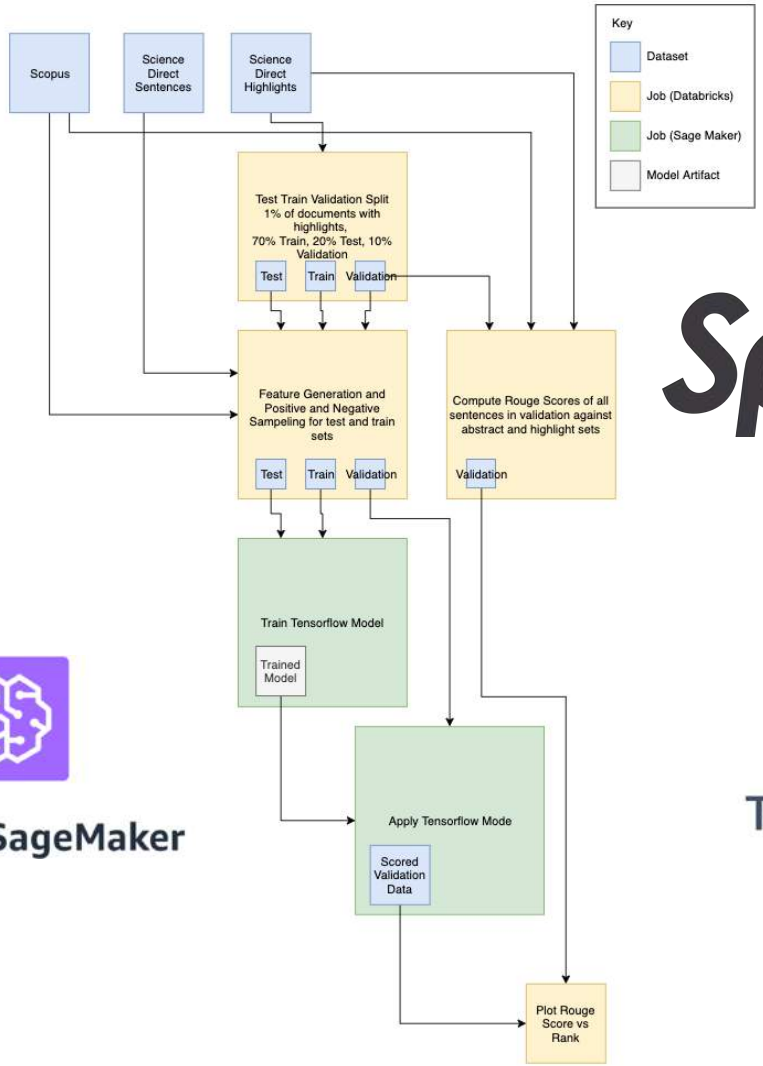


LOSS: SPARSE SOFTMAX CROSS ENTROPY

ACCURACY: BINARY ACCURACY

Training Results




Key
Dataset
Job (Databricks)
Job (Sage Maker)
Model Artifact
Scopus
Science Direct Sentences
Science Direct Highlights
Test Train Validation Split
1% of documents with highlights,
70% Train, 20% Test, 10% Validation
Test
Train
Validation
Feature Generation and Positive and Negative Sampeling for test and train sets
Test
Train
Validation
Compute Rouge Scores of all sentences in validation against abstract and highlight sets
Validation
Train Tensorflow Model
Trained Model
Apply Tensorflow Mode
Scored Validation Data
Plot Rouge Score vs Rank

Baselines

Model Name	Test Accuracy
LSTM	0.853
Abstractnet Classifier	0.718
Combined Linear Classifier	0.696
Combined MLP Classifier	0.730
Percceptron Features Abstract Vector	0.697
Single Layer NN	0.696

Offline Metrics

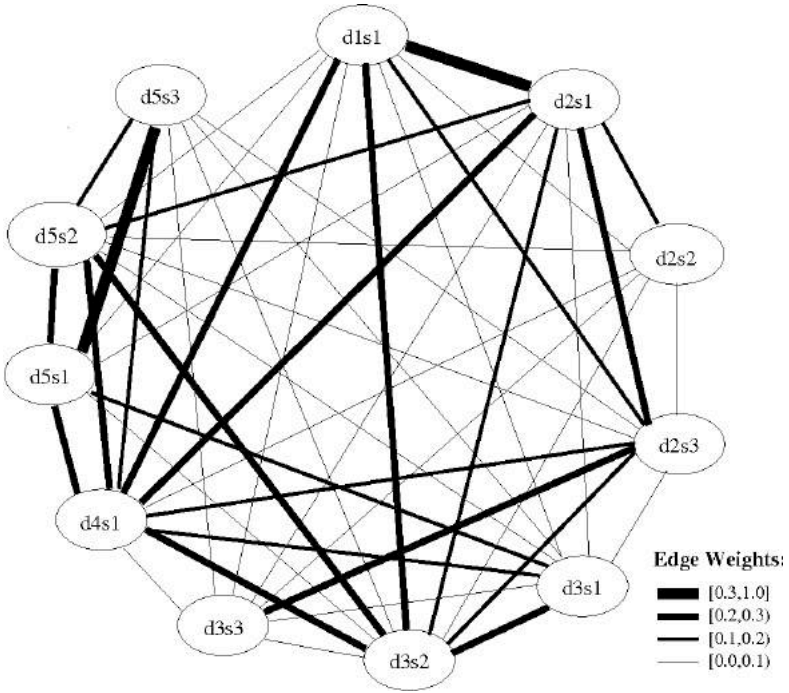
Accuracy metrics only tell one story

How well do the selected sentences compare to actual author highlights?

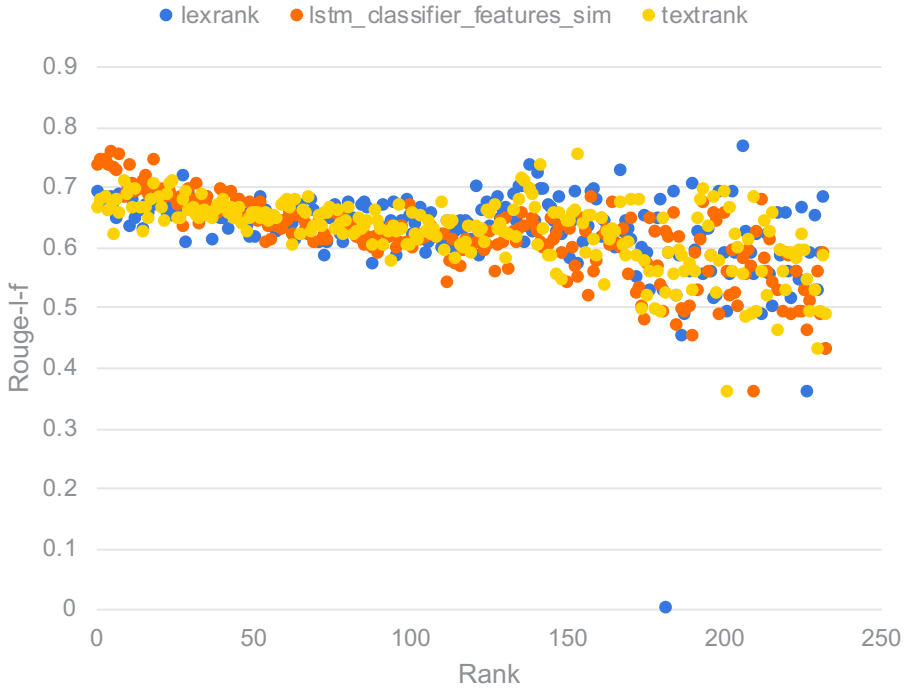
Validation set which several unseen documents, all sentences are scored and ranked

Base lines – Lex/Text Rank

- Unsupervised text summarization
- Based on page rank
- Nodes are sentences
- Edges TD-IDF between sentences
- Nodes ranked based on PageRank



Offline Metrics



	lexrank	lstm	textrank
rough@1	0.68845307	**0.73567087**	0.66500948
rough@3	0.68050251	**0.74277346**	0.68004528
rough@5	0.68086198	**0.75753316**	0.66472085
rough@10	**0.70520742**	0.68992724	0.68711934

ELSEVIER

Simplification

- Selected sentences are a tad to long.
- Contain irrelevant openings e.g. “Furthermore”
- Solution split sentences on first “,” filter out common openings.

thus
however
in summary
finally
in this study
moreover
in this work
furthermore
in addition
in conclusion
in this section
then
to the best of our knowledge
hence
in particular
additionally
also
second
first
as a result
specifically
in the present study

Simplification

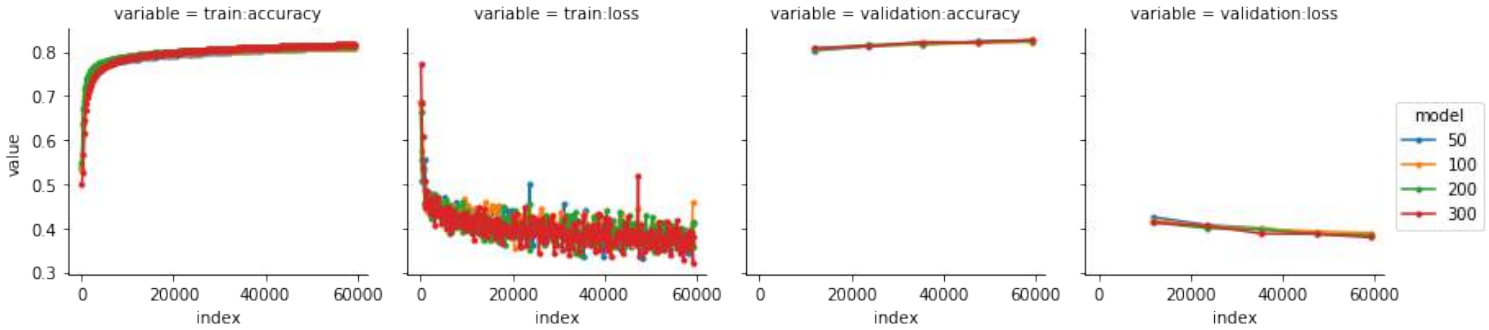
In the following work, we will design lightweight authentication protocol for three tiers wireless body area network with wearable devices.

Simplified

We will design lightweight authentication protocol for three tiers wireless body area network with wearable devices.

Effects 25% of documents

Experiments – Embedding Size



validation:accuracy 300 0.827349

In Production

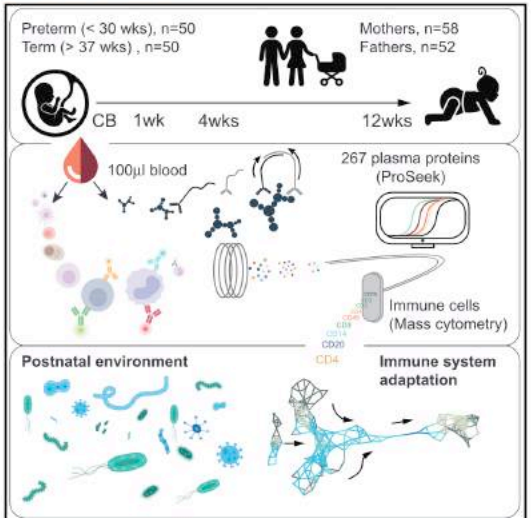
Previous PDF Next PDF

Article

Cell

Stereotypic Immune System Development in Newborn Children

Graphical Abstract



Authors

Axel Olin, Ewa Henckel, Yang Chen, ..., Cheng Zhang, Kajsa Bohlin, Petter Brodin

Correspondence

petter.brodin@ki.se

In Brief

Longitudinal profiling of blood immune cells from 100 newborns provides a systemic view on the ontogeny of the human neonatal immune system.

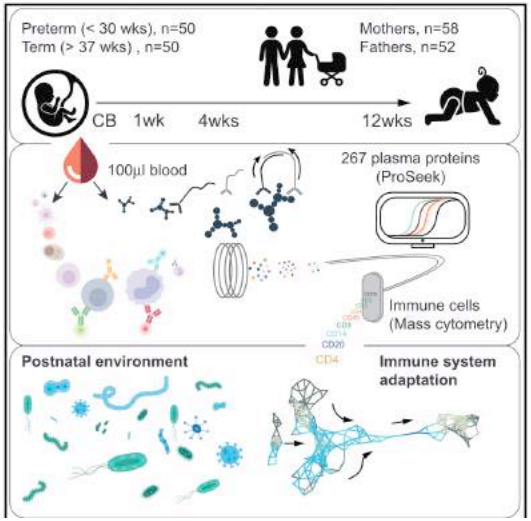
Previous PDF Next PDF

Article

Cell

Stereotypic Immune System Development in Newborn Children

Graphical Abstract



Authors

Axel Olin, Ewa Henckel, Yang Chen, ..., Cheng Zhang, Kajsa Bohlin, Petter Brodin

Correspondence

petter.brodin@ki.se

In Brief

Longitudinal profiling of blood immune cells from 100 newborns provides a systemic view on the ontogeny of the human neonatal immune system.

Highlights

- We also describe evidence of a critical period in the development of B, NK, and DCs during the first 3 months of life, as these cel...
- If microbial stimuli present during the first 100 days have similar effects on DC development, this might establish an individual's...

Show more

Click

Recommended Articles

B cell alterations during BAFF inhibition with belimumab in SLE

Daniel Ramsköld, ... +12 ... , Vivianne Malmström

EBioMedicine • February 2019

Preview View PDF Save PDF

Development and regulation of immune responses in pre- and postnatal life

Harald Renz

Clinical Biochemistry • May 2011

Preview View PDF Save PDF

Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life

Moran Yassour, ... +18 ... , Mikael Knip

Cell Host & Microbe • 11 July 2018

Preview View PDF Save PDF

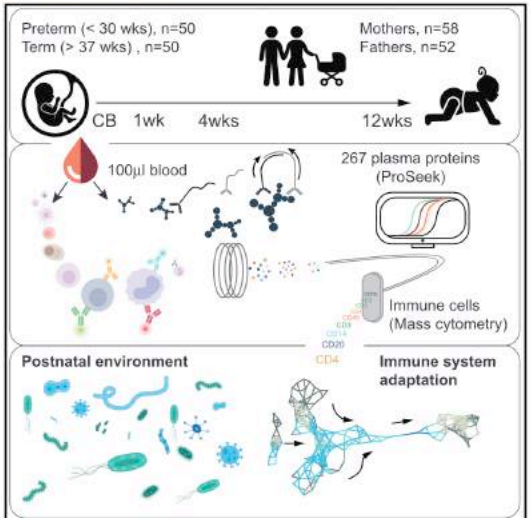
Previous PDF Next PDF

Article

Cell

Stereotypic Immune System Development in Newborn Children

Graphical Abstract



Authors

Axel Olin, Ewa Henckel, Yang Chen, ..., Cheng Zhang, Kajsa Bohlin, Petter Brodin

Correspondence

petter.brodin@ki.se

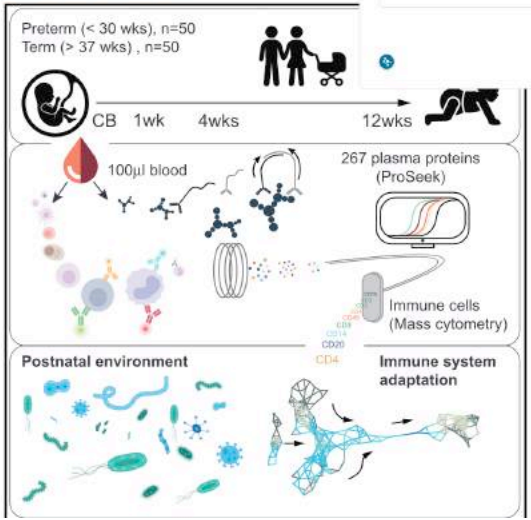
In Brief

Longitudinal profiling of blood immune cells from 100 newborns provides a systemic view on the ontogeny of the human neonatal immune system.

Cell

Stereotypic Immune … Newborn Children

Graphical Abstract



petter.brodin@ki.se

In Brief

Longitudinal profiling of blood immune cells from 100 newborns provides a systemic view on the ontogeny of the human neonatal immune system.

Article info Hide

Highlights

- We also describe evidence of a critical period in the development of B, NK, and DCs during the first 3 months of life, as these cell populations reach adult-like phenotypes during this period, suggesting that environmental influences imprinting on these cells during this time window could have long-term consequences.
- If microbial stimuli present during the first 100 days have similar effects on DC development, this might establish an individual's DCs on a trajectory associated with reduced disease risk.
- We also propose that in-depth analyses during early life adaptation to environmental influences provides a unique opportunity for better understanding the molecular mechanisms of immune system adaptation to environmental influences in humans.
- These results show that immune cell compositional changes after birth follow a stereotypic pattern of development in all children, preterm and terms alike, despite their differences in both maturity and postnatal environmental conditions.
- This also suggests that specific cell populations and pathways have different critical periods of calibration when they would be most amenable to environmental imprinting, allowing specific exposures at specific time points in the context of a given genetic makeup to contribute to an individual's risk of individual immune-mediated diseases.
- This converged 3-month immune system state might therefor represent the real set point from which human immune system variation is shaped by environmental exposures over the course of life.

Recommended Articles

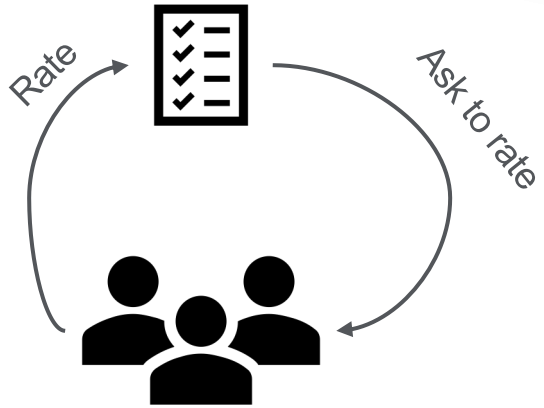
B cell alterations during BAFF inhibition with

Subject Matter Evaluation

“Human in the loop” validation framework

Work with subject matter experts (SME)

1. Ask SMEs to rate the output of the machine learning model
2. Have multiple rates rate the same output
3. Use this time help train the model



Agnostic framework, which also allows for the generation of gold standard training set for assertions

Framework used with the Lancet editors to evaluate computer generated summaries/assertions

Document being assesed

Title

Anti-inflammatory substances can influence some glial cell types but not others

Abstract

In rat microglial enriched cultures, expressing Toll-like receptor 4, we studied cytokine release after exposure with 1 ng/ml LPS for 0.5–24 h. Dexamethasone and corticosterone exposure served as controls. We focused on whether naloxone, ouabain, and bupivacaine, all agents with reported anti-inflammatory effects on astrocytes, could affect the release of TNF-α and IL-1β in microglia. Our results show that neither ultralow (10À12 M) nor high (10À6 M) concentrations of these agents had demonstrable effects on cytokine release in microglia. The results indicate that anti-inflammatory substances exert specific influences on different glial cell types. Astrocytes seem to be functional targets for anti-inflammatory substances while microglia respond directly to inflammatory stimuli and are thus more sensitive to antiinflammatory substances like corticoids. The physiological relevance might be that astrocyte dysfunction influences neuronal signalling both due to direct disturbance of astrocyte functions and in the communication within the astrocyte networks. When the signalling between astrocytes is working, then microglia produce less pro-inflammatory cytokines.

Link

10.1016/j.brainres.2013.09.052

Highlights

- Microglia are considered as immunocompetent cells of the CNS and are activated during pathological events such as stroke, ischaemia, or brain trauma to cause a neuroinflammation
- The microglial cells were stained with antibodies against the microglial specific antigen OX42, and with antibodies against TLR4, revealing that these cells do express TLR4 receptors)

What is the overall quality of the collection of highlights? (required)

	1	2	3	4	
Very Good	○	○	○	○	Very Bad

Is this collection of highlights relevent to the article (required)

	1	2	3	4	
Very Good	○	○	○	○	Very Bad

http://bit.ly/lancs-f8

ELSEVIER

Thank you

Interesting links

https://towardsdatascience.com/illustrated-guide-to-recurrent-neural-networks-79e5eb8049c9

https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21