



北京邮电大学

Beijing University of Posts and Telecommunications

深層学習に基づく学際的主題における 頻出項の発見

趙 秋涵^{1,2,3}

指導教官：楊文川 教授²

1. 東北師範大学・日本語予備学校 2班（情報・人間科学）
2. 情報工学研究科 人工知能・知能情報処理研究室
3. HP: www.zhaoqiuhan.cn

発表の流れ

- 研究の動機
- 研究方法
- 結果と結論
- 将来の研究

発表の流れ

- 研究の動機
- 研究方法
- 結果と結論
- 将来の研究

研究の動機

- これまでのデータベースまたは検索ツールの分類方法では、**学際的な主題、**いわゆる**頻出項を発見することは困難**である。

例 1 人工知能 = 「数学、統計学、コンピュータサイエンス」

例 2 [自然言語処理の発展と変遷](#) [Zhao+, 2019]

研究の動機

内容と目的

Input: **科学技術文献バッチ** (batch of literatures)

—> 深層学習 (Deep Learning)

—> 文献記述子 (descriptor)

—> 適応パラメータクラスタリング
(Adaptive Parameters Clustering, APC)

Output: **頻出項の発見と分析**

発表の流れ

- 研究の動機
- 研究方法
- 結果と結論
- 将来の研究

データ収集と前処理

深層学習

APCアルゴリズム

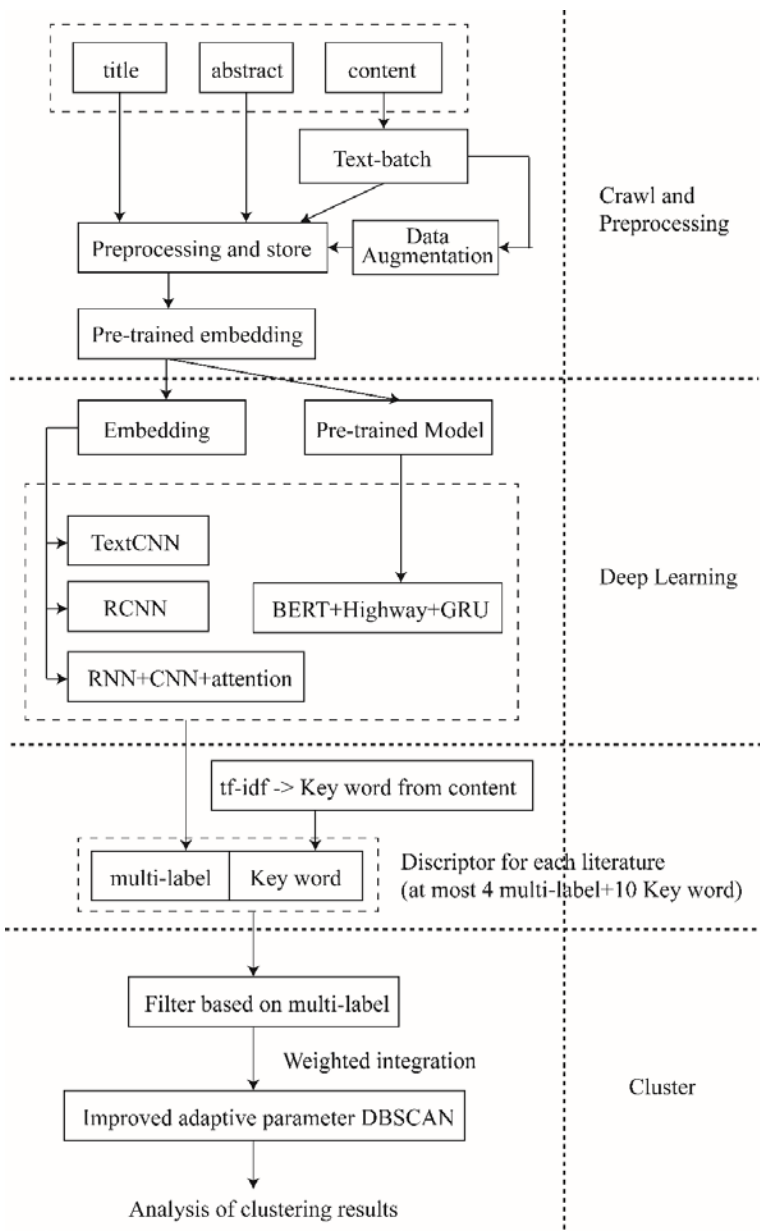
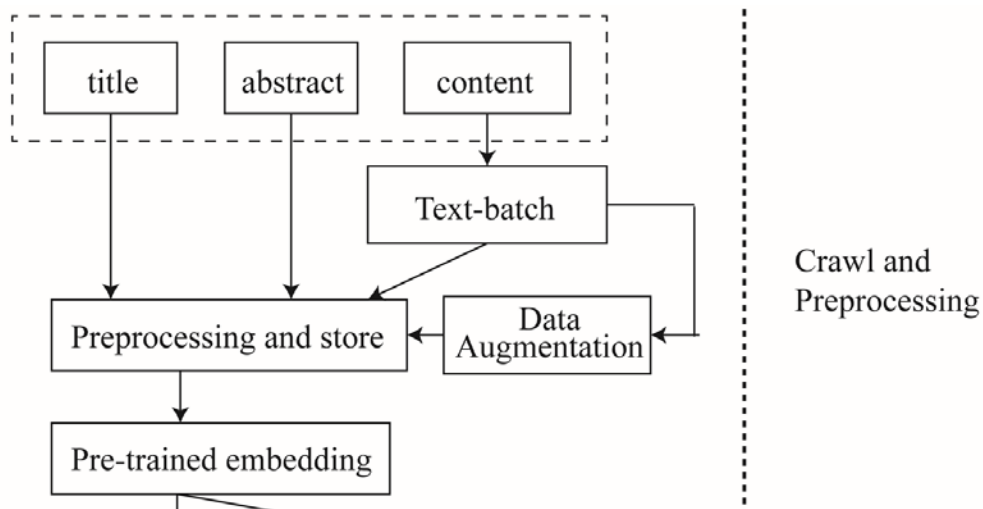


図 全体的なフレームワーク



データ収集 と前処理

Source: CNKI www.cnki.net

--> Distributed Hadoop

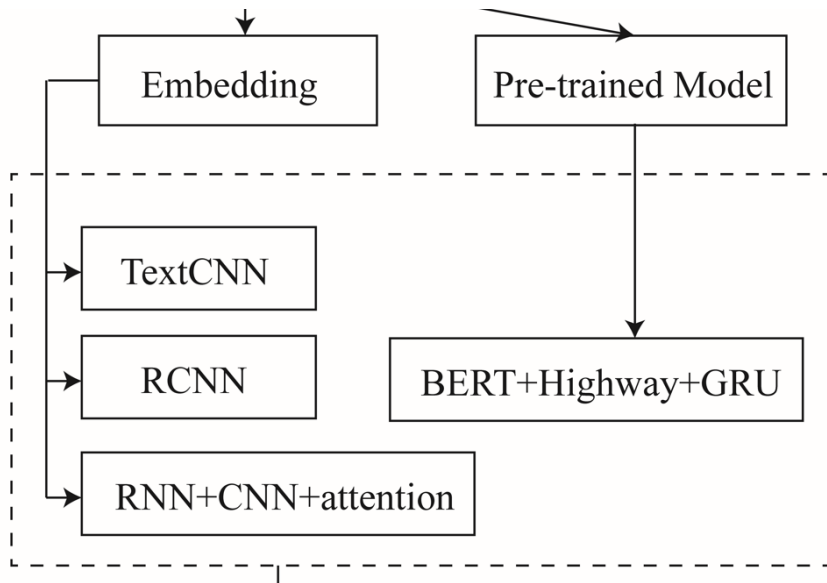
10のクラスで合計168のサブクラス
つまり $168 * 500 = 84000$ 篇科学文献

--> Word Segmentation

--> Encoding

--> Data Augmentation and Pre-training

前処理



Deep Learning

深度学习

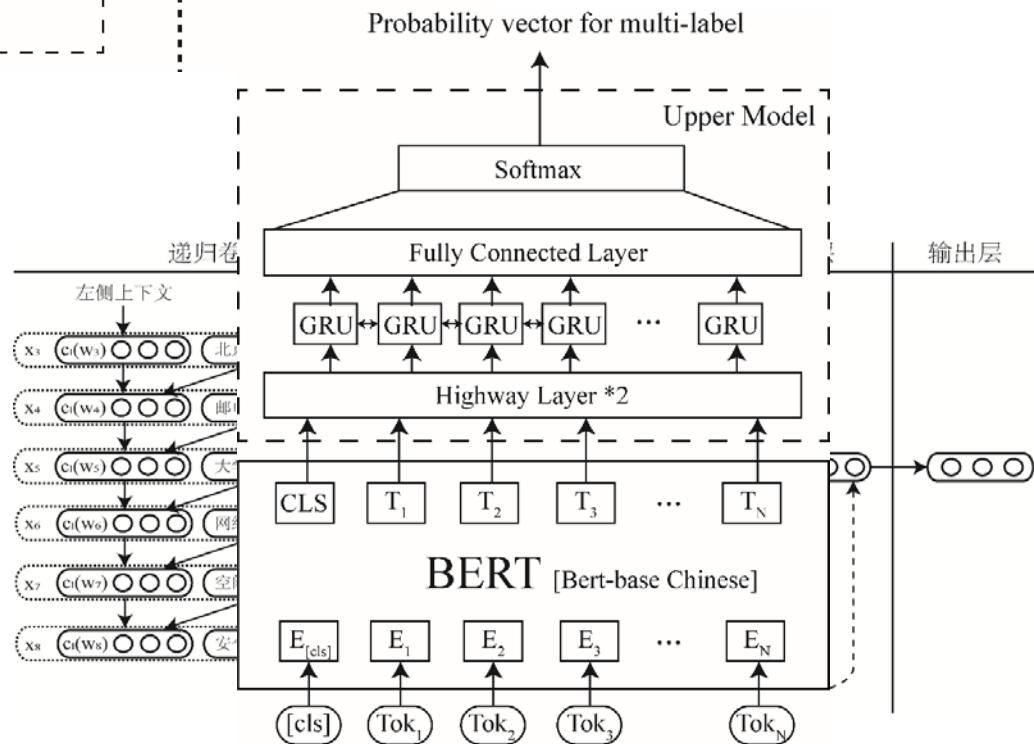
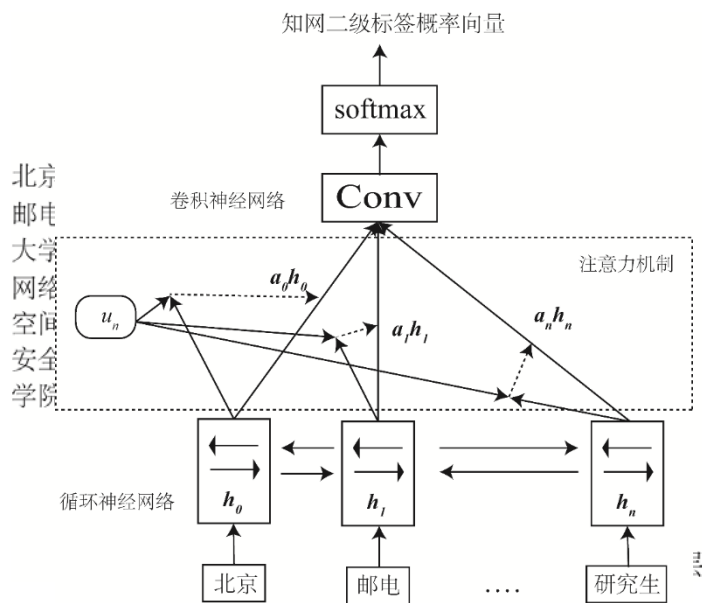
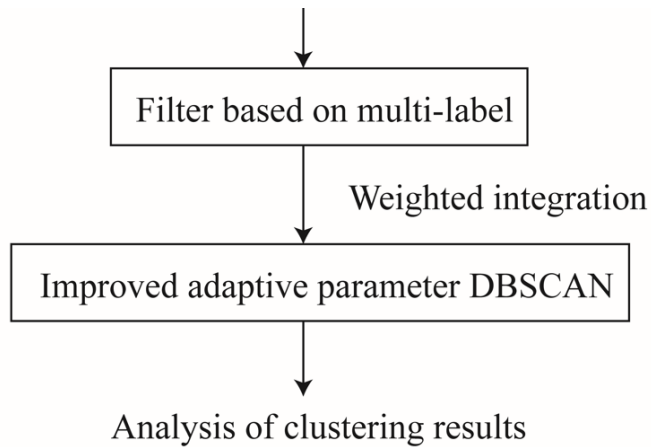


图 1 (c) RNN+CNN+TextCNN [Zhou, 2014, 2019]

图 1 (d) Bert+Highway+GRU [Zhou, 2016, 2019]



Cluster

APC¹ アルゴリズム

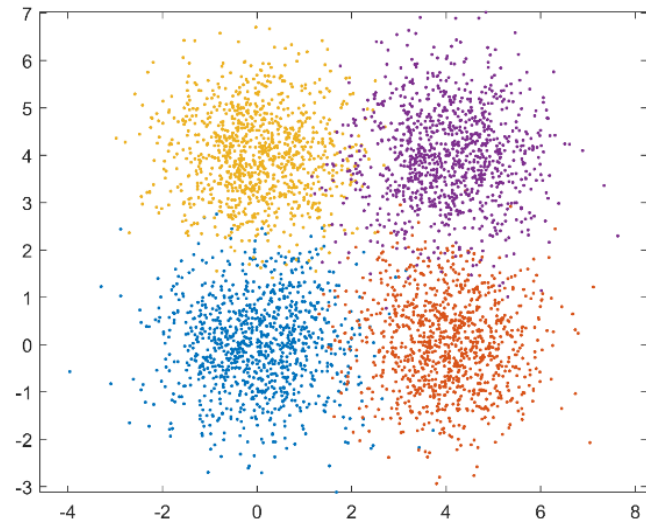
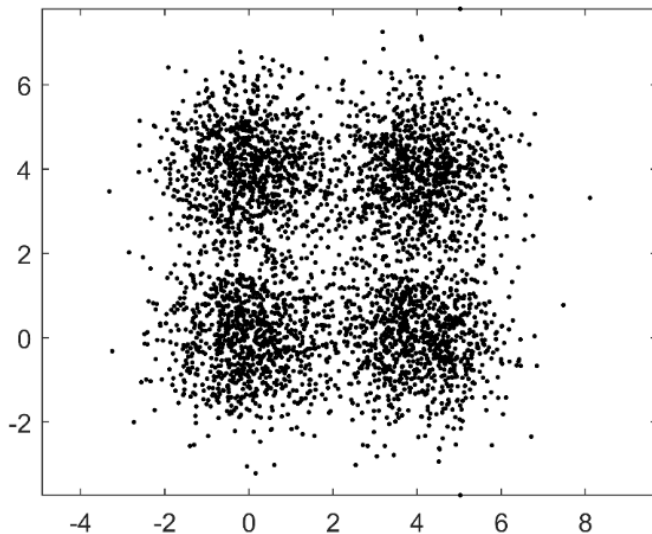


図2 クラスタリング効果の検証

1. アルゴリズムの流れは、以下のwebページに掲示している
(pp. 19) www.zhaoqiuhan.cn/files/slide/yjsbs.pdf

発表の流れ

- 研究の動機
- 研究方法
- 結果と結論
- 将来の研究

結果と結論

4つのモデルの比較

- データ拡張 (Data Augmentation)
- 事前トレーニング (pre-training)
- batch_size=32

表 1 4つのモデルの比較

アイテム	指標 $P/R/F_1$	コンバージェンス
TextCNN	0.8102 / 0.8024 / 0.8063	24
RCNN	0.8213 / 0.8160 / 0.8186	24
CNN+RNN+attention	0.8281 / 0.8377 / 0.8329	25
BERT(FT+TM)	0.8650 / 0.8555 / 0.8602	36

結果と結論

batch_sizeの最適化

- batch_size (GPU: 4 * TITAN X)
- 1 * TITAN X -> batch_size \in [8,64]

表 2 batch_sizeの最適化

batch_size	指標 F_1	コンバージョン
32	0.8602	36
64	0.8662	34
128	0.8725	33
256	0.8740	33

結果と結論

学習率の最適化

- batch_size = 256 (固定)
- 動態学習率 (automatic learning rate)

表 3 学習率の最適化

学習率	指標 F_1	コンバージェンス
0.00001	0.8736	37
0.00005	0.8740	33
0.0001	0.8709	30
動態学習率	0.8742	35

結果と結論

トレーニングの策略(training method)

- batch_size = 256 (固定)
- 動態学習率 (固定)
- FT(Fine-tune) + TM(Training-model)

表 4 トレーニング策略の影響

策略	指標 F_1	コンバージョン
NFT + TM	0.8736	37
FT + TM	0.8740	33

結果と結論

例：実際のデータへの適用

- CNKI www.cnki.net
- 2020年2月5日 “新冠 病毒” (コロナウイルス)

新聞とジャーナル合計218篇

--> 本論文の方法

--> 最大クラス10篇 (2篇を例に説明する)

結果と結論

表5 応用分析の例

文献の概要 (Google Translation)	予測のマルチレベル
青海省で新型コロナウイルスに感染した患者が、旅程を隠蔽した.. 故意に公共の場所に侵入したり、他の人との接触を隠したりする人は犯罪を構成すること.. [Yang, 2020]	ウイルス学 法律学
「新型コロナウイルス」の発見と蔓延に伴い.. 労働者と雇用主の労使関係はどのように.. 感染症の予防と管理に関する法律第3条に従って.. [Yan, 2020]	ウイルス学 法律学

発表の流れ

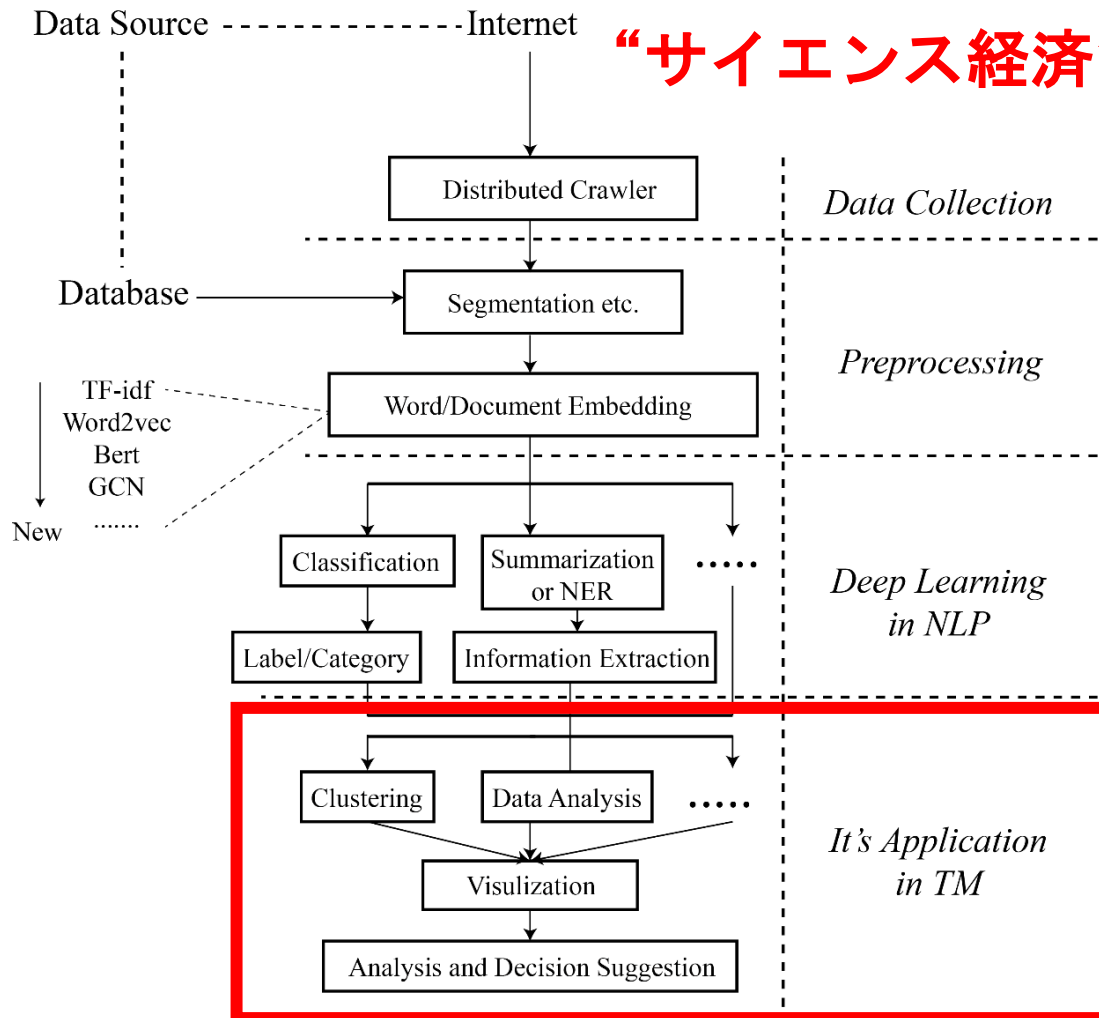
- 研究の動機
- 研究方法
- 結果と結論
- 将来の研究

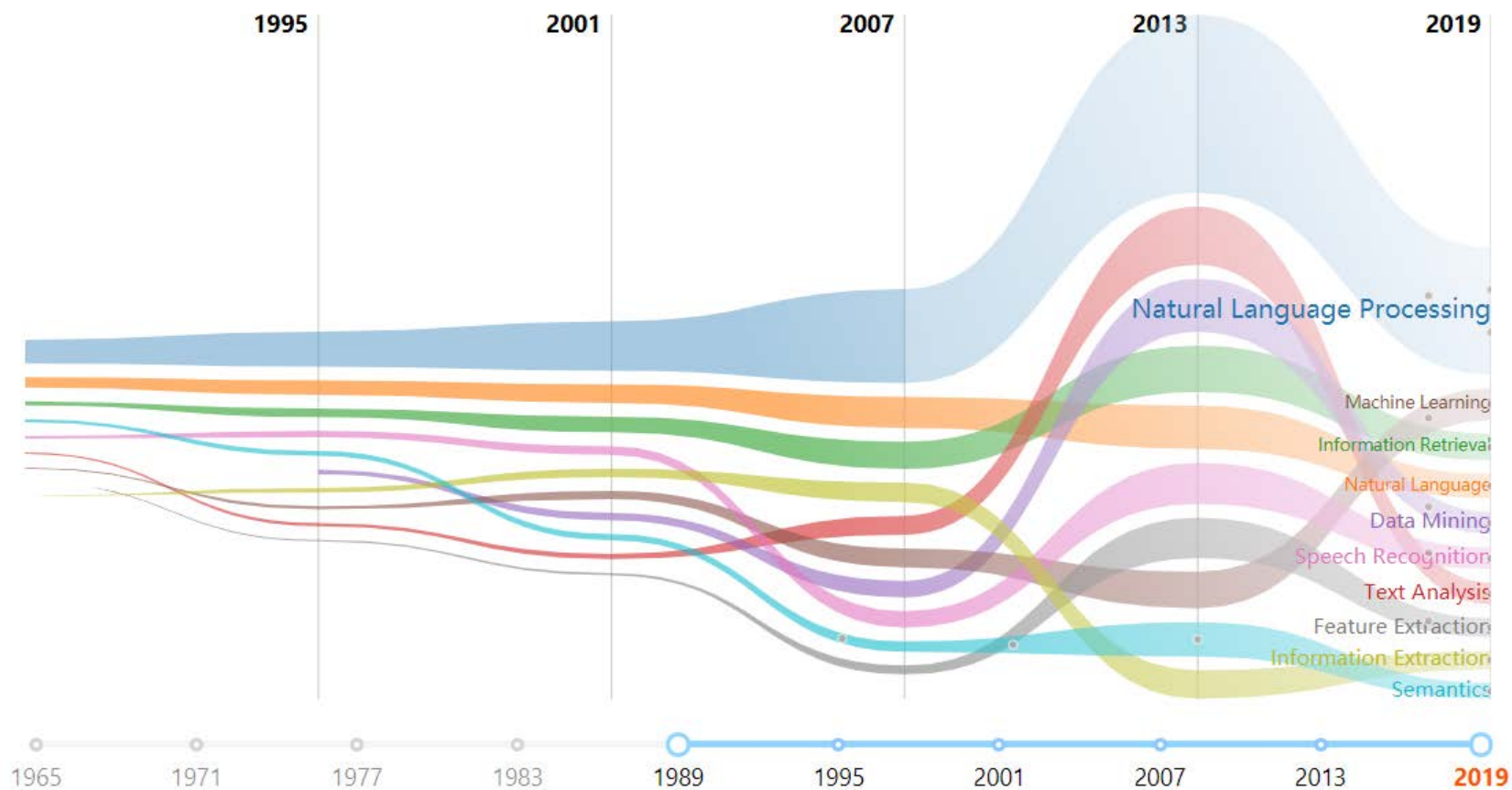
将来の研究

- 博士課程は東京大学の工学研究科で、元橋一之先生のもと、技術経営戦略学（TMI）についての研究を進めていく。
- 具体的には、中米貿易戦を背景に、自然言語処理と深層学習を組み合わせて、科学文献と特許データを通じて、中国とアメリカの革新的な産業（AI、Bigdata、IoTなど）の動向と発展を分析することである。

将来の研究

“サイエンス経済” への応用





自然言語処理

- > (1989年) 「自然言語、情報検索、語彙分析」
- > (2007年) 「テキスト分析、データマイニング、情報検索」
- > (2019年) 「機械学習、自然言語、情報検索」