



「次世代人工知能・ロボット中核技術開発」
(人工知能分野) 中間成果発表会
－人間と相互理解できる人工知能に向けて－

自然言語理解を核とした データ・知識融合技術の研究開発

平成29年3月29日

国立情報学研究所

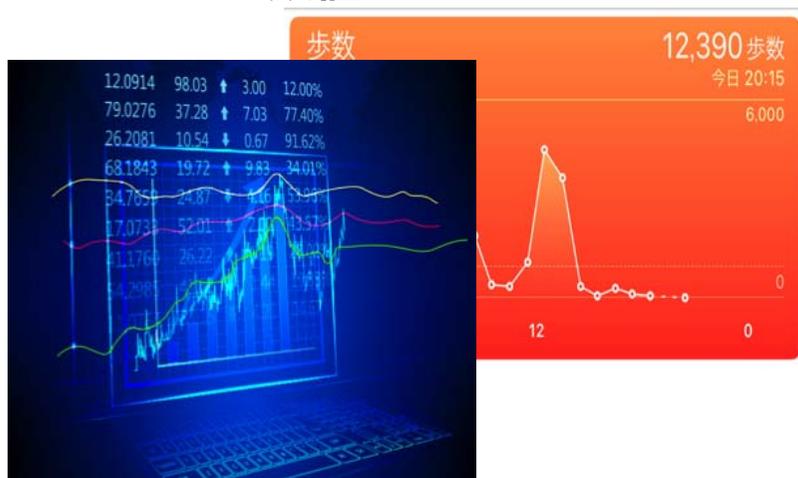
宮尾 祐介

国立研究開発法人 産業技術総合研究所

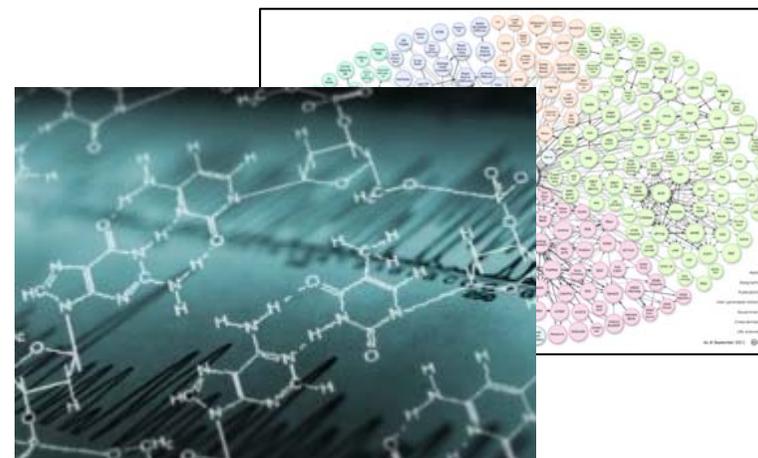
国立研究開発法人 新エネルギー・産業技術総合開発機構

世の中のデータは、さまざまな形をしている

数値データ



データベース・オントロジー



画像・映像データ



A screenshot of a Wikipedia article for 'Janissary'. The article includes a definition, a list of contents, and an image of a Janissary soldier. The contents list includes: 1 Origin, 2 Characteristics, 3 Recruitment, training and status, 4 Training, 5 Janissary corps, 6 Corps strength, 7 Disbandment, 8 Status, 9 See also, and 10 Notes and references.

自然言語データ

研究開発の目的

・ 自然言語を通して多様なデータを理解する技術

数値データ



長期金利が0.1%上がった時、IT関連株にはどのような影響があるか？

データベース・オントロジー

2000年以降に設立された学部で、定員割れしているところをリストアップ



子供が公園で遊んでいる場面が見たい

画像・映像データ



Janissary

From Wikipedia, the free encyclopedia

For the Janissary series of novels by Jerry Brownlee, see *Janissary series*.

The **janissaries** (from Ottoman Turkish *çelebi* *çelebi* meaning "free soldier"; *Adnan*: *Jeniper*, *Donner*: *Jambet*, *Crusader*: *Jarghat*, *Pomorian*: *Jenest*, *Serbian*: *Jarghat*) were military Muslim slave units that formed the Ottoman sultan's household troops and bodyguards. The first was created by the Sultan Murad I from Christian boys seized through the *devşirme* system from conquered countries in the 14th century^[1] and was abolished by Sultan Mahmud I in 1626 in the *Asupozu* incident^[2].

Contents [hide]

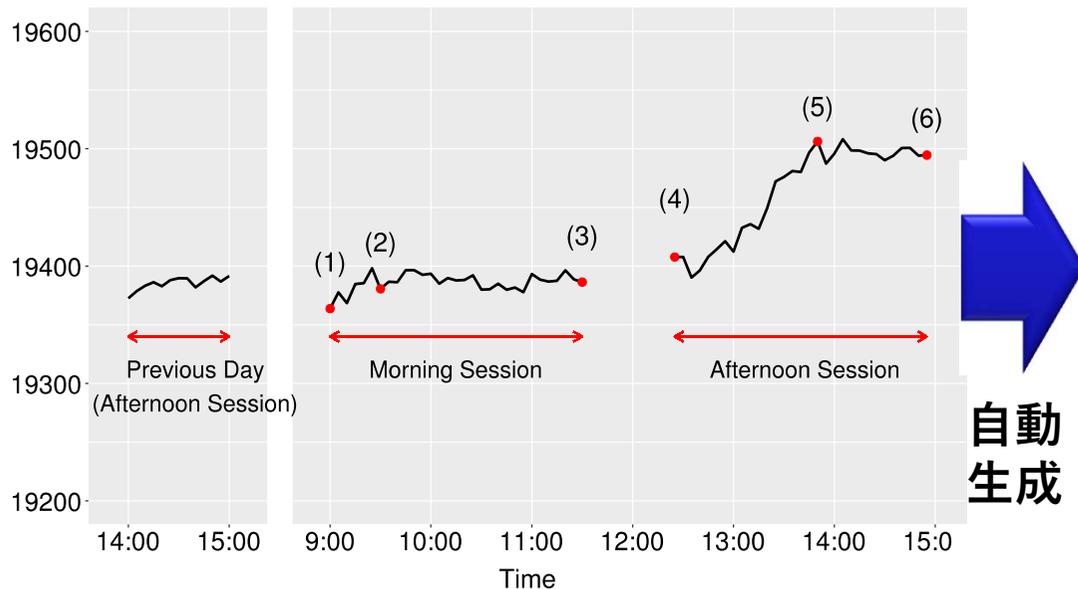
- Origin
- Characteristics
- Recruitment, training and status
- Training
- Janissary corps
- Corps strength
- Commanders
- Salaries
- Abolition and reestablishment

自然言語データ

時系列数値データ

- 株式市場、気象、センサー、etc.
- データから何をどのように読み取るか？
- 株式市場の概況を自動生成する問題を例に

時系列数値データ

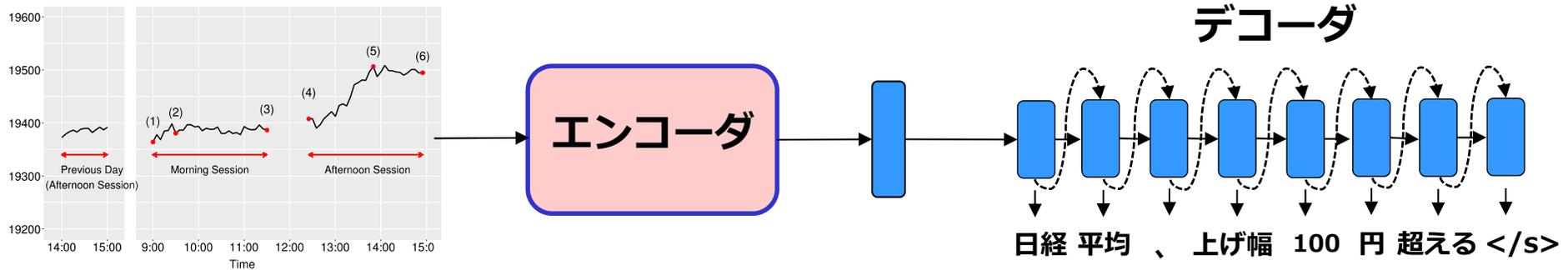


概況テキスト

- | | | |
|-----|-------|---|
| (1) | 09:00 | 日経平均、 続落 で始まる |
| (2) | 09:29 | 日経平均、 上げ に転じる |
| (3) | 11:30 | 日経平均、 続落 前引けは 5円安 の 19386円 |
| (4) | 12:30 | 日経平均、 午後 は 上昇 で始まる |
| (5) | 13:54 | 日経平均、 上げ幅100円 を超える |
| (6) | 15:00 | 日経平均、 反発 大引けは 102円高 の 19494円 |

よくある手法： エンコーダ・デコーダモデル

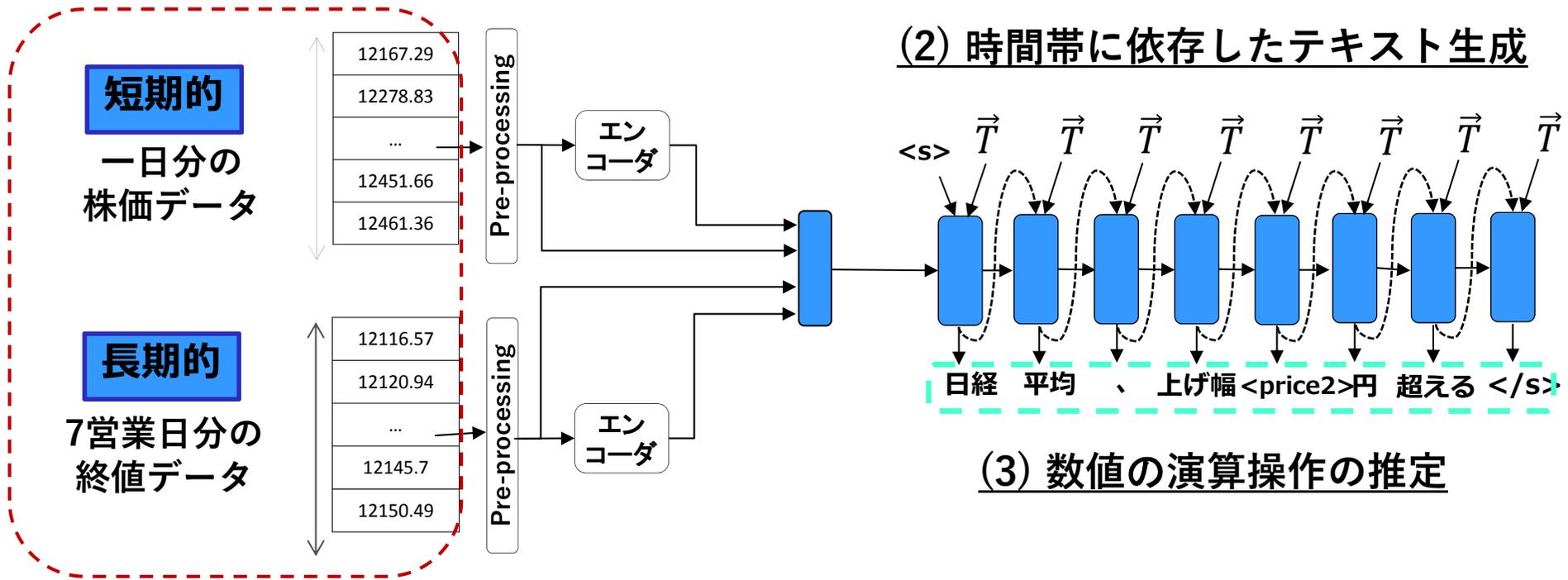
- 入力を実数ベクトルにエンコードし、そこから文章を生成する
- 最近、自然言語生成で広く用いられている
 - 機械翻訳、文書要約、画像説明文生成、etc.
- 学習データ（入出力のペア）が大量に与えられれば、自動文章生成が高精度で実現できる
- 既存手法では、限られたデータからは概況テキストの多様性を学習できない



正解	日経平均、反落 大引けは 29 円 安の 16735 円
エンコーダ・デコーダモデル	日経平均、 続伸 前引けは <unk> 円 高 の 19937 円

[村上ら、2017]

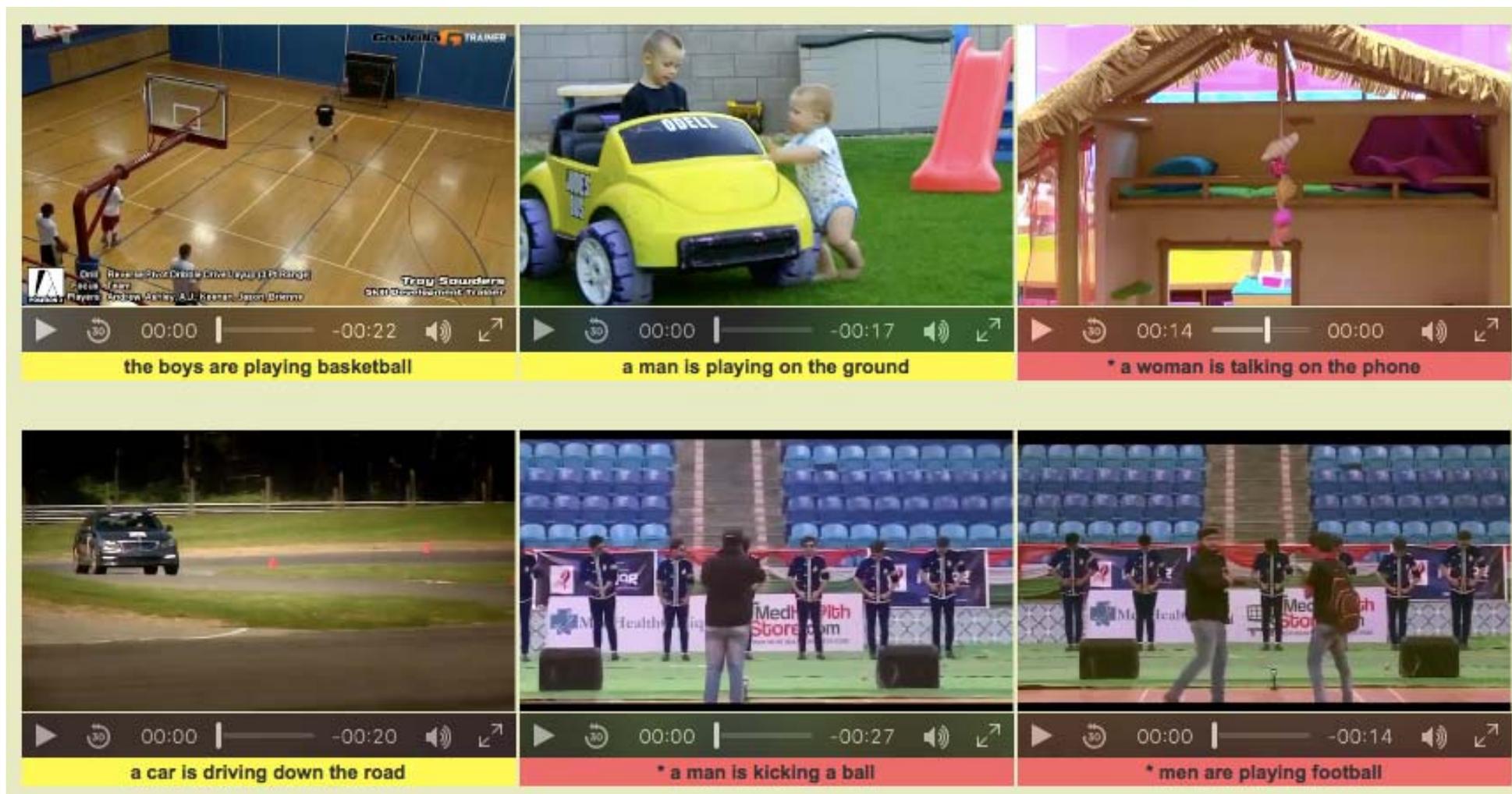
- 時系列数値データの特徴をとらえ、テキスト化する手法を提案
- 数千文の学習データで、ほぼ完璧な文を生成



(1) 時系列数値データの短期的・長期的特徴のエンコード

[Natsuda et al. 2016]

- 映像データに対する説明文生成
 - 映像中の重要なフレームに注目するモデル



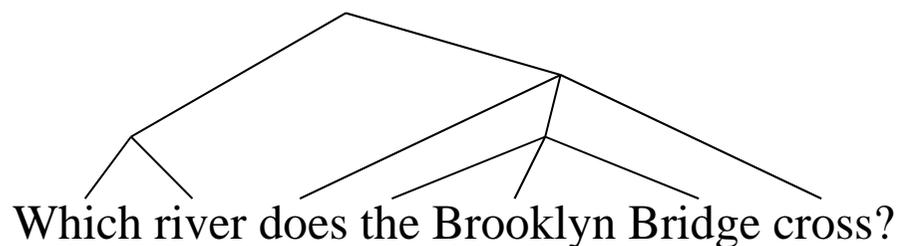
[Martinez-Gomez et al. 2016]

- さまざまな分野の知識が、大規模データベースに蓄積される
 - Freebase, DBPedia, 生命科学データベース、etc.
 - データベースの知識を有効活用したい
- 自然言語の質問文をデータベースクエリに変換
- 木構造変換とノード変換の最適な組合せを探索

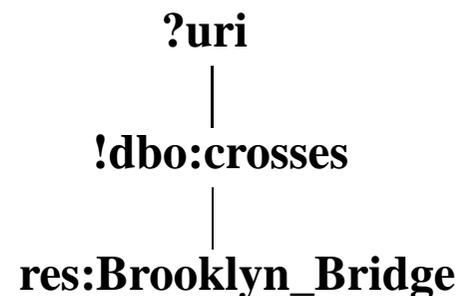
データベースクエリ

質問文

Which river does the Brooklyn Bridge cross?



```
SELECT DISTINCT ?uri WHERE {
  res:Brooklyn_Bridge dbo:crosses ?uri .
}
```



おわりに

- さまざまなデータを自然言語を通して理解する
 - 時系列数値データ
 - 画像・映像データ
 - 構造化データベース
- 今後の展望
 - さまざまなデータについて、自然言語テキストとおなじようにアクセスが可能となる
 - 検索、質問応答、要約、etc.
 - マルチモーダル対話、ロボットの視覚・言語情報理解など、様々な応用が期待される