Transformerを用いた非構造データの構造化

2020/11/20

人工知能学会合同研究会　知識ベースシステム研究会
株式会社リーディング・エッジ社　研究開発部　川崎拳人

概要

・弊社リーディング・エッジ社では、IT人材の派遣事業を主にやっており、
求人情報を含むメール文書を扱う機会が多い

・しかしながら、このようなメール文書はフォーマットがバラバラであり、
営業戦略において、価値あるデータとして扱うことが難しい状況にある

・そこで、近年注目されてきている、深層学習や自然言語処理技術を用いて、
非構造化データである求人情報を含むメール文書の構造化を行った

目次

非構造化データとは

非構造化データは構造化データに比べ、AIやデータサイエンスの分野では生かし切れていない

	非構造化データ	構造化データ
特徴	・フォーマットが決まっていない ・データ量が多い分、ストレージコストがかかる ・データの検索や更新に手間がかかる ・活用するためには、データに前処理を施す必要がある	・フォーマットが決まっている ・管理しやすい ・検索や更新に手間がかからない ・行列で管理されているのですぐに活用できる状態にある
例	文書、画像、音声など	DB、CSVなど

本研究の目的

弊社に送られてくる案件メールをJSONに変換する

表1 メール文

【案件名】 AI Webアプリ開発案件
【内容】 チャットボットを用いたWebアプリの設計・開発に携わっていただきます
【勤務地】 茅場町 （東京メトロ東西線・日比谷線）
【期間】 10月～12月 ※延長の可能性あり
【必須スキル】 Python, C#を使用した開発経験 自然言語処理に関する知識
【推奨スキル】 javascript, django Webアプリ開発の経験
【単価】 50～60万前後
【年齢】 40歳はじめまでを希望
【人数】 1名
【面談回数】 2回（弊社同席）
【その他】 勤怠，コミュニケーションに問題のない方

表2 JSON

{"job_name": ["AI Webアプリ開発案件"],
"contents": ["チャットボットを用いたWebアプリの設計・開発に携わっていただきます"],
"sites_station": ["茅場町"],
"durings": ["10月～12月"],
"price_min": [500000],
"price_max": [600000],
"age_min": [-1],
"age_max": [43],
"skill_required_os": [],
"skill_required_lang": ["Python", "C#"],
…,
"skill_recommended_web": ["JavaScript", "Django"],
"required_numbers": ["1名"],
"counts_for_interview": ["2回（弊社同席）"],
"etc": ["勤怠，コミュニケーションに問題のない方"]}

構造化における課題

メールは大量にあり、そのフォーマットもバラバラ

メール１

【案　件】大規模システム開発
医療向け、損保向け、システム開発系向け
各業種それぞれ大規模システム開発経験のある方を募集しております。
【場　所】新宿
【期　間】7月～長期
【人　数】複数名
【単　価】スキル見合い
【精　算】140-180h
【スキル】
＜必須＞
・Java3年以上の開発経験
・JavaScript1年以上の開発経験
＜尚可＞
・大規模システム開発経験者
・ネットワーク経験者
【面　談】2回
【外国籍】不可
【備　考】コントロール問題ない方

メール２

案 件 名：Kubernetes環境の基盤構築
作業概要：
・インフラ・運用・自動化
・WEBサーバの基盤構築
必須スキル：
・Linuxに関して構築経験3年以上
・サーバ基盤の設計構築経験3年以上
歓迎スキル：
・Python
・Chef、Jenkins
・Zabbix
・Kubernetesを使用した経験
期間：即日～長期　※3月からでも可能
場所：渋谷
募集：1～2名
単価：65万～80万（スキル見合い）
精算：あり
面談：1回（場合によっては2回）
年齢：制限なし
時間：9:00～18:00

メール３

◆作業内容：外資企業向け会計システムのエンハンス業務
・調査、問合せ依頼に対して、システムの調査
・既存システムの改修に伴う、設計、オフショアへの展開、受入れ検証
・本番リリース
◆現場環境：UNIX、Oracle、COBOL、JCL
◆必須：UNIXの操作経験
K-shellの改修経験
◆尚可：COBOL経験
◆作業場所：大井町駅徒歩5分/JR高崎駅徒歩15分
◆期間：即日～長期
◆勤務時間　9：30～18：30
◆予算：Max50万（140-200）
◆人数：1名（40歳までを希望）
◆面談：Max2回（1回目弊社同席）

構造化における課題

構造化するためには、様々な処理が必要

メール１

【案　件】大規模システム開発
医療向け、損保向け、システム開発系向け
各業種それぞれ大規模システム開発経験の
ある方を募集しております。
【場　所】新宿
【期　間】7月～長期
【人　数】複数名
【単　価】スキル見合い
【精　算】140-180h
【スキル】
＜必須＞
・Java3年以上の開発経験
・JavaScript1年以上の開発経験
＜尚可＞
・大規模システム開発経験者
・ネットワーク経験者
【面　談】2回
【外国籍】不可
【備　考】コントロール問題ない方

項目名の判定
抜き出す要素の判定
スキル要素の抽出
可/不可をフラグで表す

メール２

案 件 名：Kubernetes環境の基盤構築
作業概要：
・インフラ・運用・自動化
・WEBサーバの基盤構築
必須スキル：
・Linuxに関して構築経験3年以上
・サーバ基盤の設計構築経験3年以上
歓迎スキル：
・Python
・Chef、Jenkins
・Zabbix
・Kubernetesを使用した経験
期間：即日～長期　※3月からでも可能
場所：渋谷
募集：1～2名
単価：65万～80万（スキル見合い）
精算：あり
面談：1回（場合によっては2回）
年齢：制限なし
時間：9:00～18:00

工程の抽出

メール３

◆作業内容：外資企業向け会計システム
のエンハンス業務
・調査、問合せ依頼に対して、システム
の調査
・既存シ　　う、設計、
オフショ　　入れ検証
・本番リリース
◆現場環境：UNIX、Oracle、COBOL、
JCL
◆必須：UNIXの操作経験
K-shellの改修経験
◆尚可：COBOL経験
◆作業場所：大井町駅徒歩5分/JR高崎駅
徒歩15分
◆期間：即日～長期
◆勤務時間　9：30～18：30
◆予算：Max50万（140-200）
◆人数：1名（40歳までを希望）
◆面談：Max2回（1　　自弊社同席）

工程の抽出
最寄り駅の抽出
項目と対になっていない要素の抽出

構造化における課題

構造化するためには、様々な処理が必要

メール１

【案　件】大規模システム開発
医療向け、損保向け、システム開発系向け
各業種それぞれ大規模システム開発経験の
ある方を募集しております。
【場　所】新宿
【期　間】7月～長期
【人　数】複数名
【単　価】スキル見合い
【精　算】140-180h
【スキル】
＜必須＞
・Java3年以上の開発経験
・JavaScript1年以上の開発経験
＜尚可＞
・大規模システム開発経験者
・ネットワーク経験者
【面　談】2回
【外国籍】不可
【備　考】コントロール問題ない方

定量化する

スキルを抽出して
カテゴリ別に
振り分ける

可/不可をフラグで表す

メール２

様々な表現を統一する

案 件 名：Kubernetes環境の基盤構築
作業概要：
・インフラ・運用・自動化
・WEBサーバの基盤構築
必須スキル：
・Linuxに関して構築経験3年以上
・サーバ基盤の設計構築経験3年以上
歓迎スキル：
・Python
・Chef、Jenkins
・Zabbix
・Kubernetesを使用した経験
期間：即日～長期　※3月からでも可能
場所：渋谷
募集：1～2名
単価：65万～80万（スキル見合い）
精算：あり
面談：1回（場合によっては2回）
年齢：制限なし
時間：9:00～1

数値に変換する

定量化する

メール３

◆作業内容：外資企業向け会計システム
のエンハンス業務
・調査、問合せ依頼に対して、システム
の調査
・既存システムの改修に伴う、設計、
オフショアへの展開、受入れ検証
・本番リリース
◆現場環境：UNIX、Oracle、COBOL、
の操作経験
K-shellの改修経験
◆尚可：COBOL経験
◆作業場所：大井町駅徒歩5分/JR高崎駅
徒歩15分
◆期間：即日～長期
◆勤務時間　9：30～18：30
◆予算：Max50万（140-200）
◆人数：1名（40歳までを希望）
◆面談：Max2回（1回目弊社同席）

大文字小文字を揃える

構造化における課題

課題を構造化タスクに集約し、それを実現することによって課題を解決する

課題	フォーマットがバラバラであるため、様々な処理が必要

↓

構造化タスク	1. 特定情報の抽出 2. 表記変換 3. 表現の統一

↓

課題解決	構造化タスクの実現

構造化タスク　1. 特定情報の抽出

抽出すべき項目名とその内容の関連性を判断して，特定個所を抽出する

表1 メール文

【案件名】 AI Webアプリ開発案件

【勤務地】 茅場町 （東京メトロ東西線・日比谷線）

【期間】 10月～12月 ※延長の可能性あり

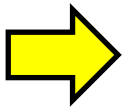


表2 JSON

各項目に合った内容を抽出

"job_name": ["AI Webアプリ開発案件"],

"sites_station": ["茅場町"],

"durings": ["10月～12月"],

構造化タスク　2. 表記変換

同一の意味を持つ様々な表記を統一する

表1 メール文

【単価】 50～60万前後

【推奨スキル】
javascript, django Webアプリ開発の経験

表2 JSON

上限値/下限値を判断して数値に変換する

"price_min": [500000],
"price_max": [600000],

値の文字列をパスカルケースに変換する

"skill_recommended_web":
["JavaScript", "Django"]

構造化タスク　3. 表現の統一

抽象的な表現を定量的な表現に統一する

表1 メール文

【年齢】 40歳はじめまでを希望

「【年齢】40歳はじめまでを希望」は，何歳までのことを指しているのだろうか？

曖昧な表現

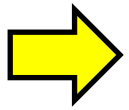


表2 JSON

"age_min": [-1],
"age_max": [43],

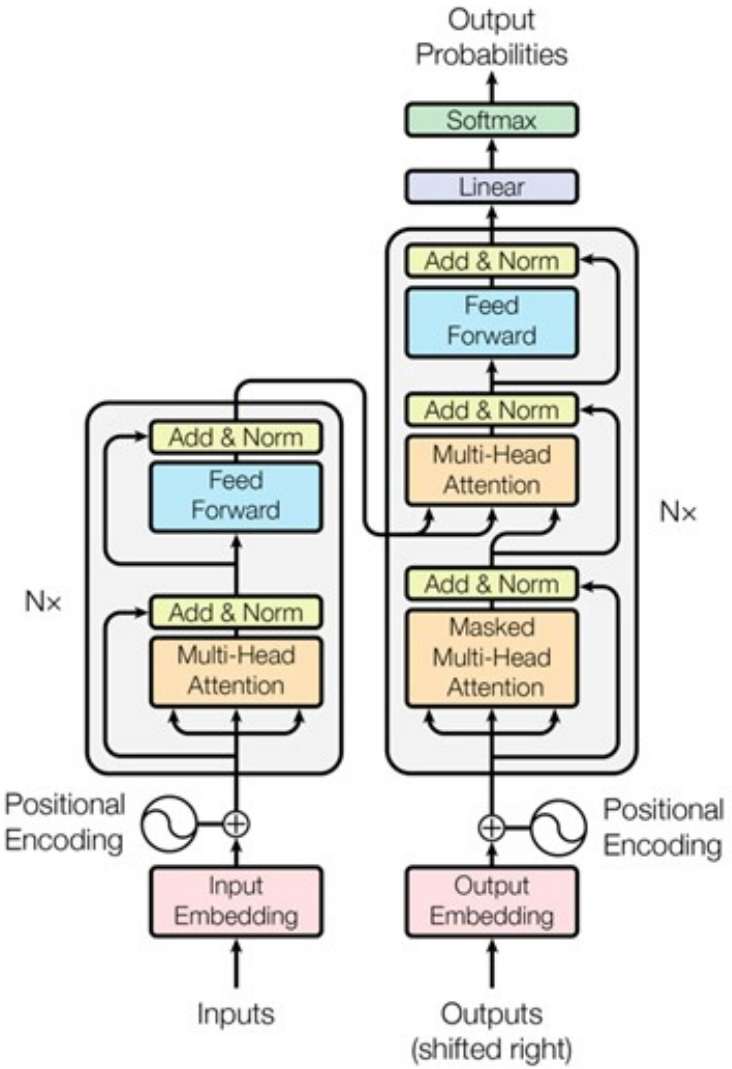
何歳～何歳までか具体的な数値で表す*

＊ [-1]は下限値に制約がないことを示す

構造化手法

・非構造化データから構造化データに“翻訳”という形で構造化を行う

Transformer



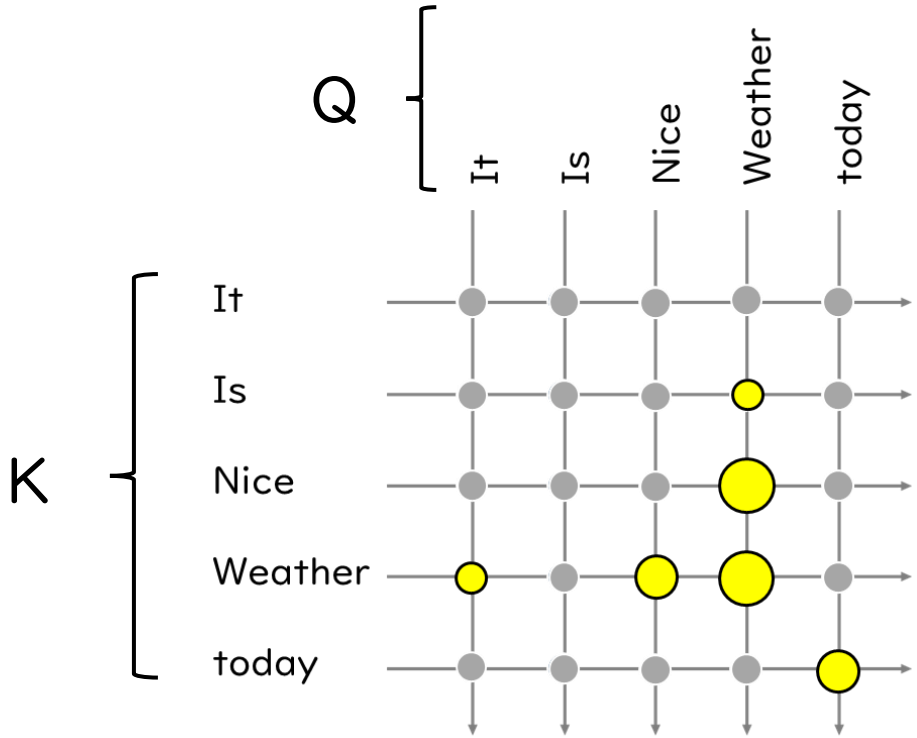
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention Is All You Need. arXiv:1706.03762(2017)

- 2017年にGoogleが論文を発表
  「Attention Is All You Need」
- 機械翻訳タスクにおいて、公開当時最先端であったSeq2seqやGNMTよりも高いスコアをマークした
- EncoderとDecoderから構成される
- 内部にAttentionメカニズムを持つ

Attention

・下図のSelf Attentionは「文章内における単語間の関連性」を学習する
・「It is nice weather today.」という文におけるAttention

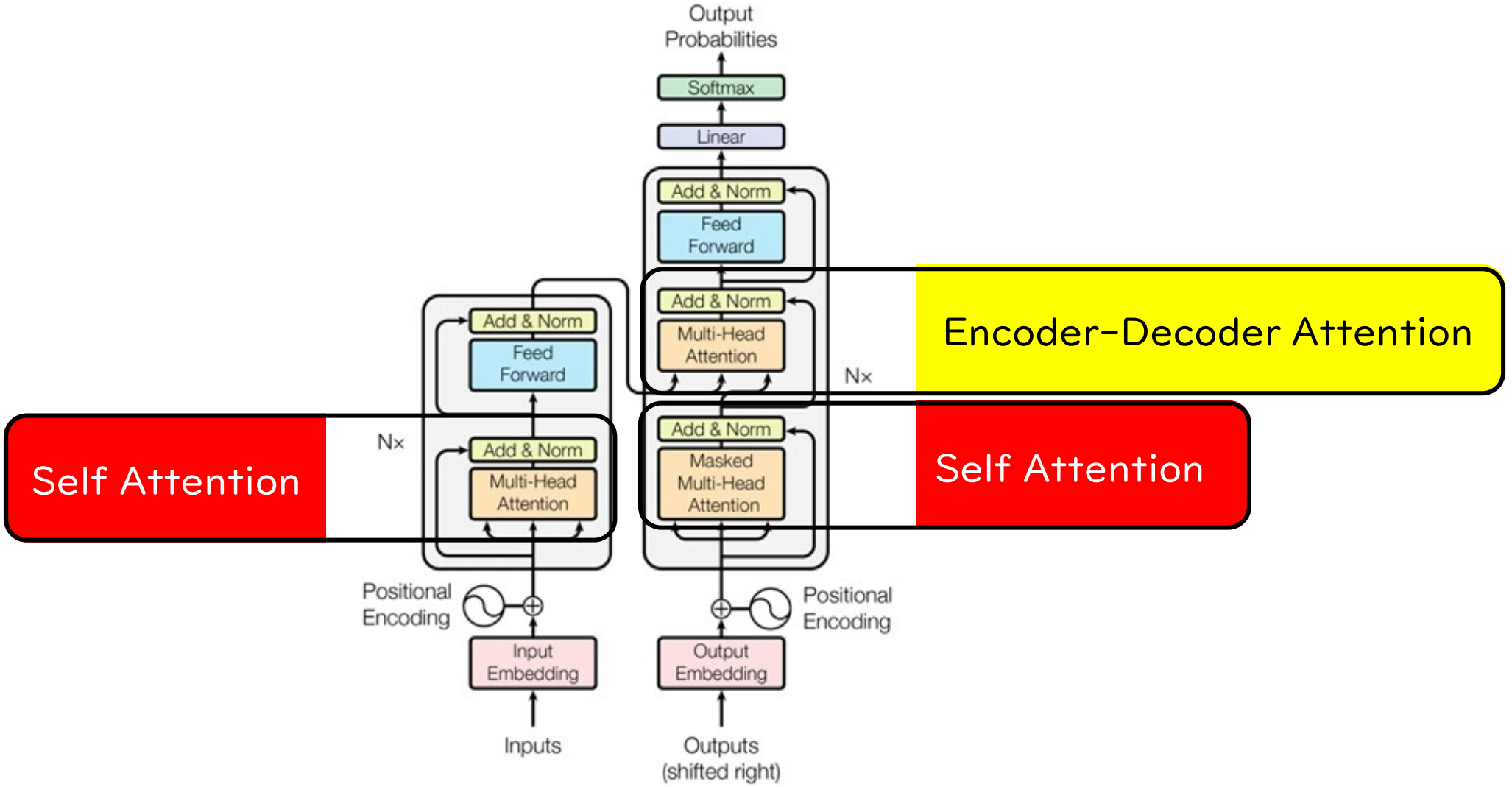
$$Attention = softmax(\frac{QK^T}{\sqrt{d_k}})V$$



✖ It is nice weather today .

softmaxにより単語の重要度が0～1の間で表される

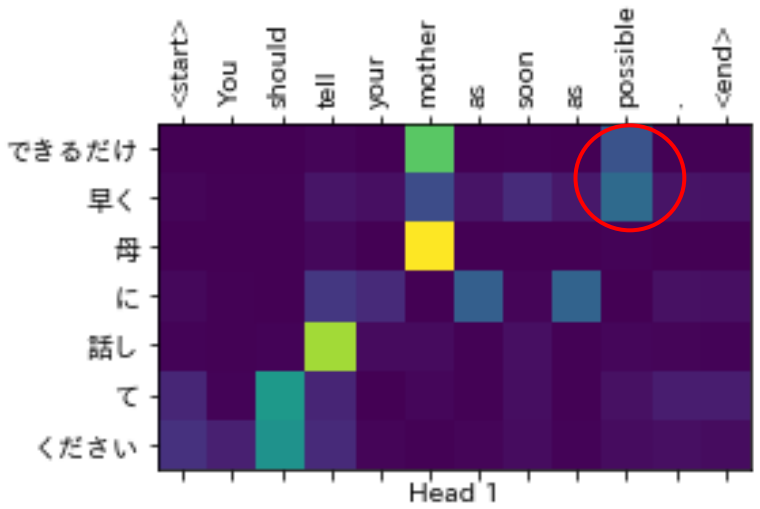
Transformer

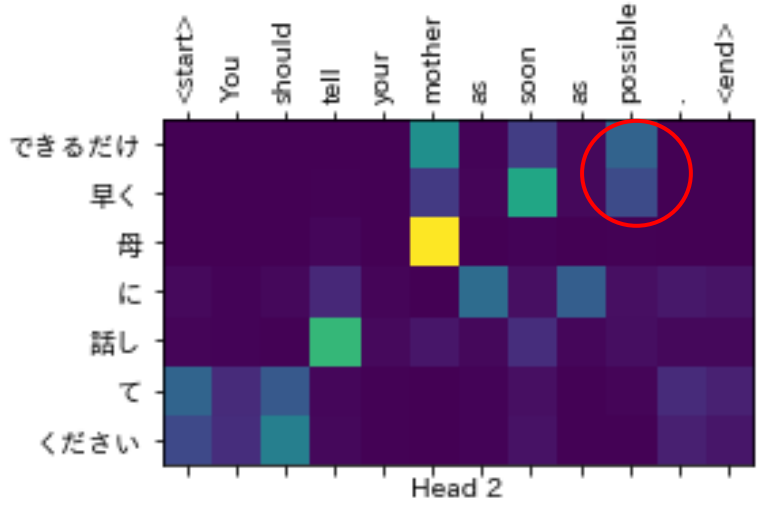


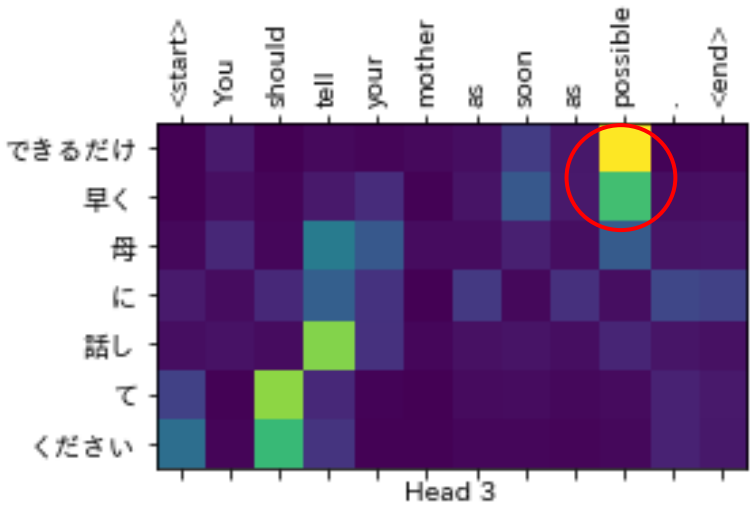
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention Is All You Need. arXiv:1706.03762(2017)

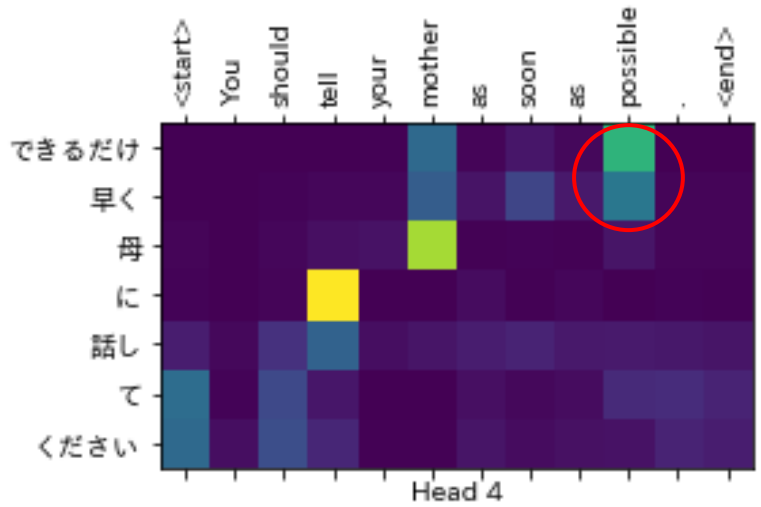
Multi Head Attention

・Attentionを複数にすることで、各々のAttentionが異なる位置の異なる部分に注意を向け，単語間の関連性を処理できるということが実験的にわかっている









Transformerを採用した理由

・Self Attention
　→　文章内の単語間の関連性から重要な情報とそうでない情報を判断する

・Encoder-Decoder Attention
　→　ソース文の単語とターゲット文の単語間の関連性を求め，
　　多種多様な表現を統一するのに有効である

・End-to- End学習
　→　ルールベースでは対応しきれない様々なケースに対応できることが期待される

実験

ソース文

表1 メール文

【案件名】 AI Webアプリ開発案件
【内容】 チャットボットを用いたWebアプリの設計・開発に携わっていただきます
【勤務地】 茅場町 （東京メトロ東西線・日比谷線）
【期間】 10月～12月 ※延長の可能性あり
【必須スキル】 Python, C#を使用した開発経験 自然言語処理に関する知識
【推奨スキル】 javascript, django Webアプリ開発の経験
【単価】 50～60万前後
【年齢】 40歳はじめまでを希望
【人数】 1名
【面談回数】 2回（弊社同席）
【その他】 勤怠，コミュニケーションに問題のない方

ターゲット文

表2 JSON

{"job_name": ["AI Webアプリ開発案件"],
"contents": ["チャットボットを用いたWebアプリの設計・開発に携わっていただきます"],
"sites_station": ["茅場町"],
"durings": ["10月～12月"],
"price_min": [500000],
"price_max": [600000],
"age_min": [-1],
"age_max": [43],
"skill_required_os": [],
"skill_required_lang": ["Python", "C#"],
…,
"skill_recommended_web": ["JavaScript", "Django"],
"required_numbers": ["1名"],
"counts_for_interview": ["2回（弊社同席）"],
"etc": ["勤怠，コミュニケーションに問題のない方"]}

実験

JSON項目一覧

Key	詳細
job_name	案件名
industry	業界
contents	案件概要
sites_station	最寄り駅
durings	期間
price_min	単価 下限値（ない場合は-1）
price_max	単価 上限値
price_skill_matchmaking	スキル見合いフラグ
age_min	年齢 下限値（ない場合は-1）
age_max	年齢 上限値
skill_required_os	必須スキル OS
skill_required_lang	必須スキル 言語
skill_required_web	必須スキル Web
skill_required_tech	必須スキル その他
skill_recommended_os	推奨スキル OS
skill_recommended_lang	推奨スキル 言語
skill_recommended_web	推奨スキル Web
skill_recommended_tech	推奨スキル その他

Key	詳細
environment_os	開発環境 OS
environment_lang	開発環境 言語
environment_web	開発環境 Web
environment_tech	開発環境 その他
required_numbers	要求人数
counts_for_interview	面談回数
working_time_start	始業時間
working_time_end	終業時間
average_work_hours	平均労働時間
settlement_to_overhours	稼働時間
organization	組織名
payment_site	支払いサイト
commercial_flow	商流
can_accept_foreigner_flag	外国人可否
can_accept_foreigner_negotiable	外国人交渉可否
etc	その他

実験

・前処理

- メール文書，JSONはどちらもテキストに変換，MeCabを用いて単語に分割した．
- JSONにおける単語分割に際しては，項目名（Key）はそれ自体が最小単位である一単語として扱った．

＜JSONの単語分割例＞

```
{ “ job_name “ : [ “ AI Web アプリ 開発 案件 ” ],
“ contents ” : [ “チャット ボット を 用いた Web アプリ の 設計 ・ 開発 に
携わ って いただき ます “ ],
“ sites_station ”:[ “茅場町 “ ],
“ durings “: [ “ 10 月 ～ 12 月 “ ],
“ price_min ": [ 500000 ],
“ age_min ": [- 1 ],
“ age_max “: [ 43 ],
“ skill_required_os ": [ ],
“ skill_recommended_web “: [ “ JavaScript “, “ Django “ ],
```

求人メールに該当するJSONの項目がない場合

値が複数ある場合

実験

- 使用データ
  - データ件数：約48,000（80%：学習用，10%：検証用，10%：テスト用）
  - 語彙数：
    - メール文書：35,201
    - JSON：24,165

■メール文書 シーケンス長の統計量

データセット	平均	中央値	標準偏差
学習	231.6	198.0	128.5
検証	231.0	199.0	127.8
テスト	229.0	197.0	125.3

■JSON シーケンス長の統計量

データセット	平均	中央値	標準偏差
学習	253.4	238.0	70.4
検証	254.0	238.0	70.8
テスト	252.7	237.0	69.8

評価方法と結果

・データ件数：約4,800

・評価方法：予測によって得られたJSONと正解コーパスのJSONにおいて，
同じ項目の値をExact MatchとF値で評価

・結果：

Exact Match	F値
90.44	96.59

ケーススタディ　特定情報の抽出

◇案件名　.某大手通信会社向け…

↓

job_name:[某大手通信会社向け…]

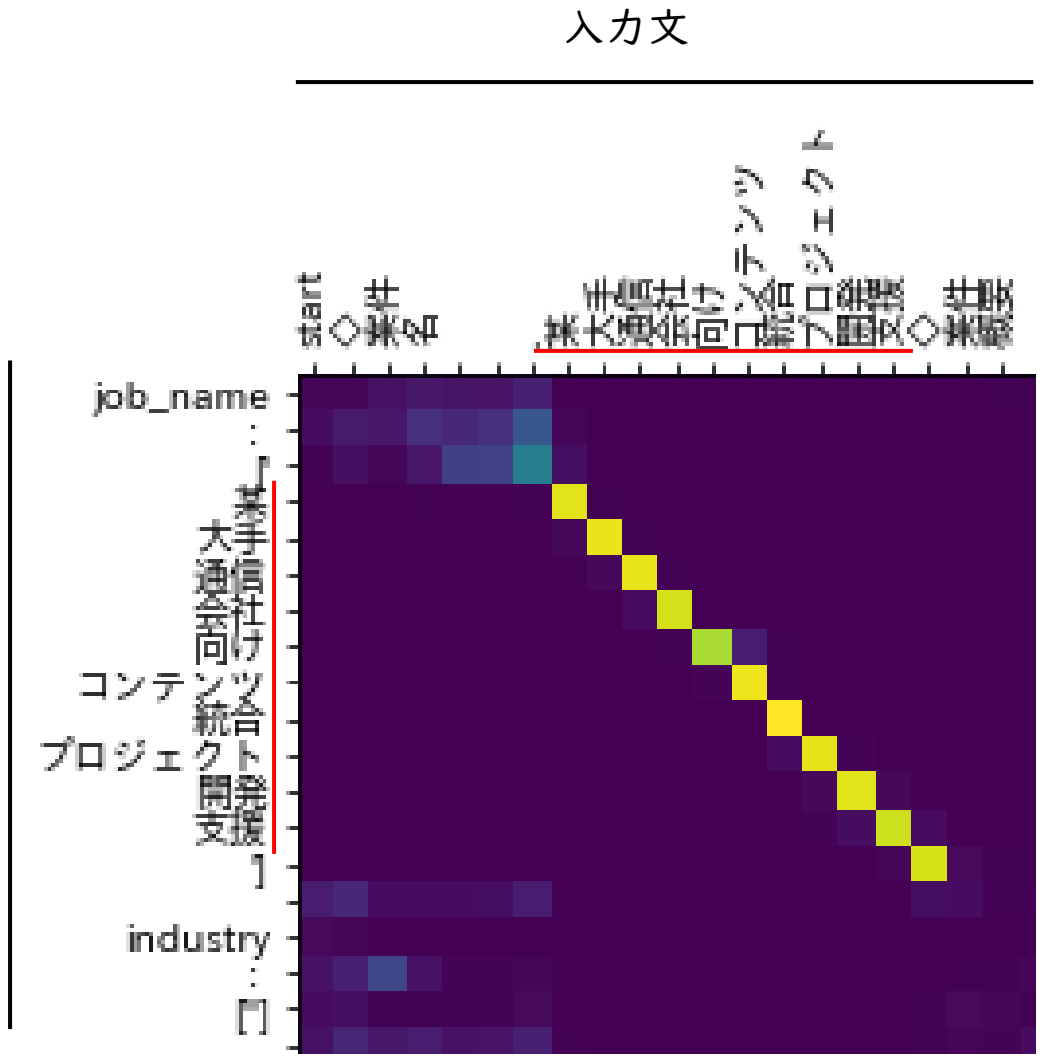


図１　job_nameにおけるAttention Weight

ケーススタディ

表記変換

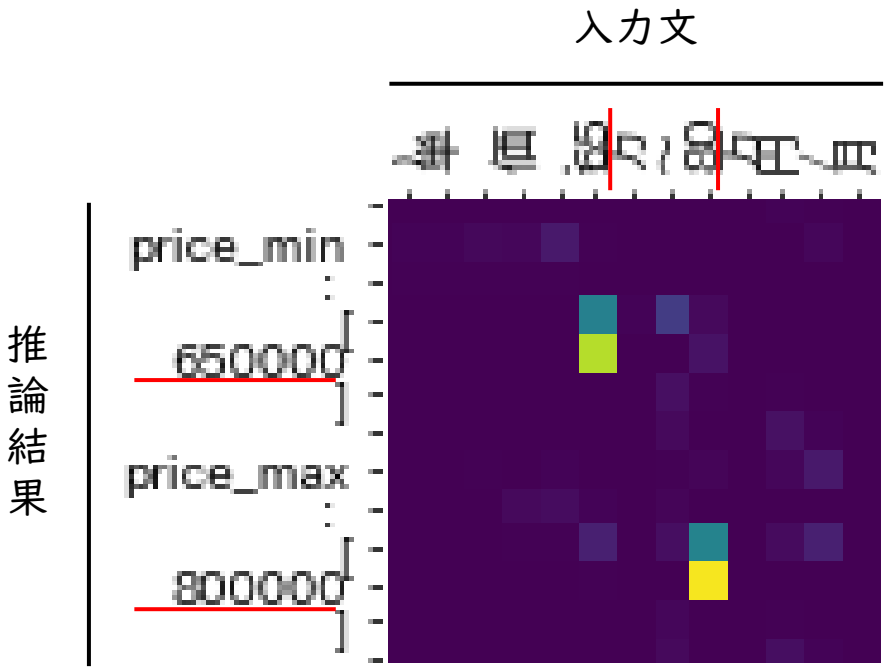


図2　priceにおけるAttention Weight

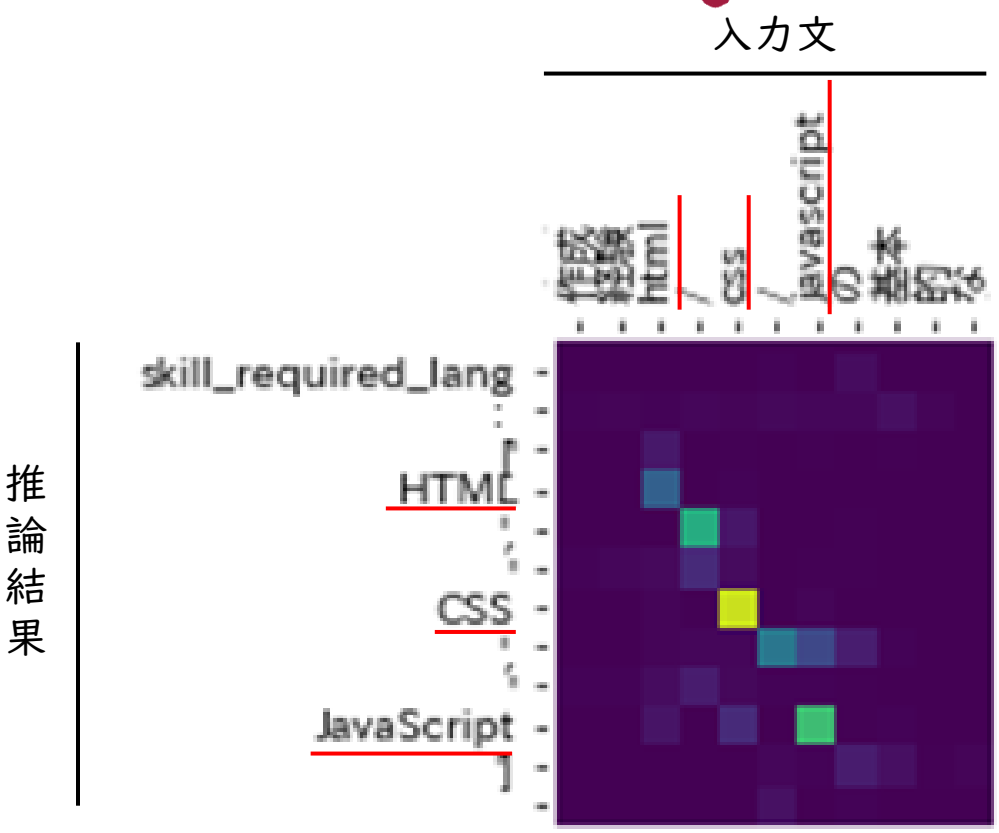


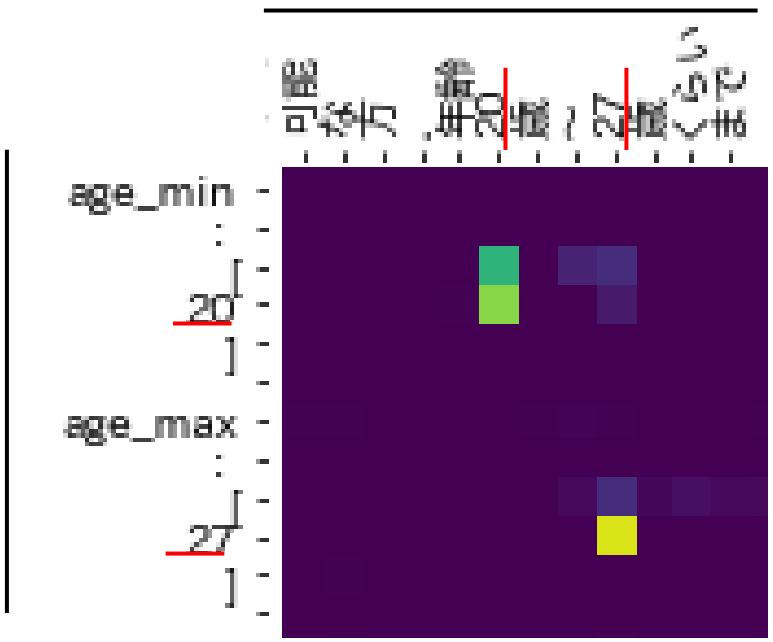
図3　skill_required_langにおけるAttention Weight

単価. 65万～80万円/月
↓
price_min:[650000]
price_max:[800000]

作成経験html/css/javascriptの基本的な
↓
Skill_required_lang:[‘HTML’ , ’ CSS’ , ‘JavaScript’]

ケーススタディ　表現の統一

入力文

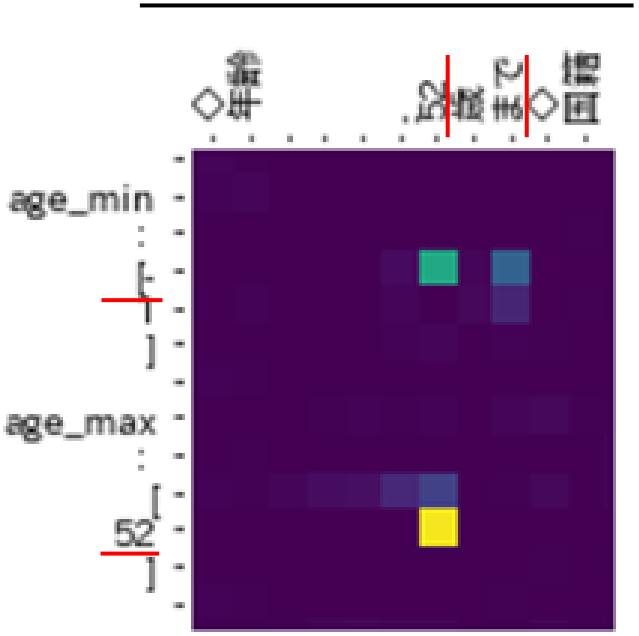


推論結果

図4　ageにおけるAttention Weight

年齢20歳～27歳くらいまで
↓
age_min:[20]
age_max:[27]

入力文



推論結果

図6　ageにおけるAttention Weight

◇年齢　.52歳まで◇国籍
↓
age_min:[-1]
age_max:[52]

まとめ

・非構造化データであるメール文書の構造化を行った

・構造化の手法は機械翻訳に用いられるTransformerを採用し、
　３つの構造化タスク（特定情報の抽出、表記変換、表現の統一）に取り組んだ

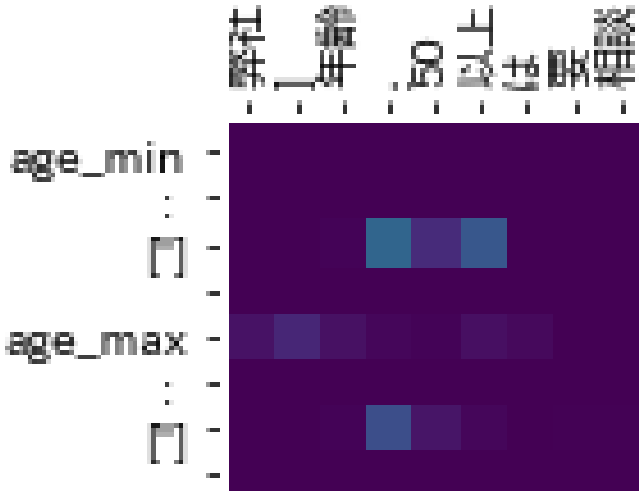
・特定情報の抽出や表記変換では、Transformerの有効性が示された

・一方、課題として表現の統一が不十分であることや一部誤変換が確認された

課題

1. 表現の統一

複数の単語の連なりによって意味が形成される場合、うまく構造化できない



年齢. 50以上…　→　“age_min”: [50]　？

後に続く単語によって文脈が変化する

年齢. 50以上は要相談　→　“age_max": [50]

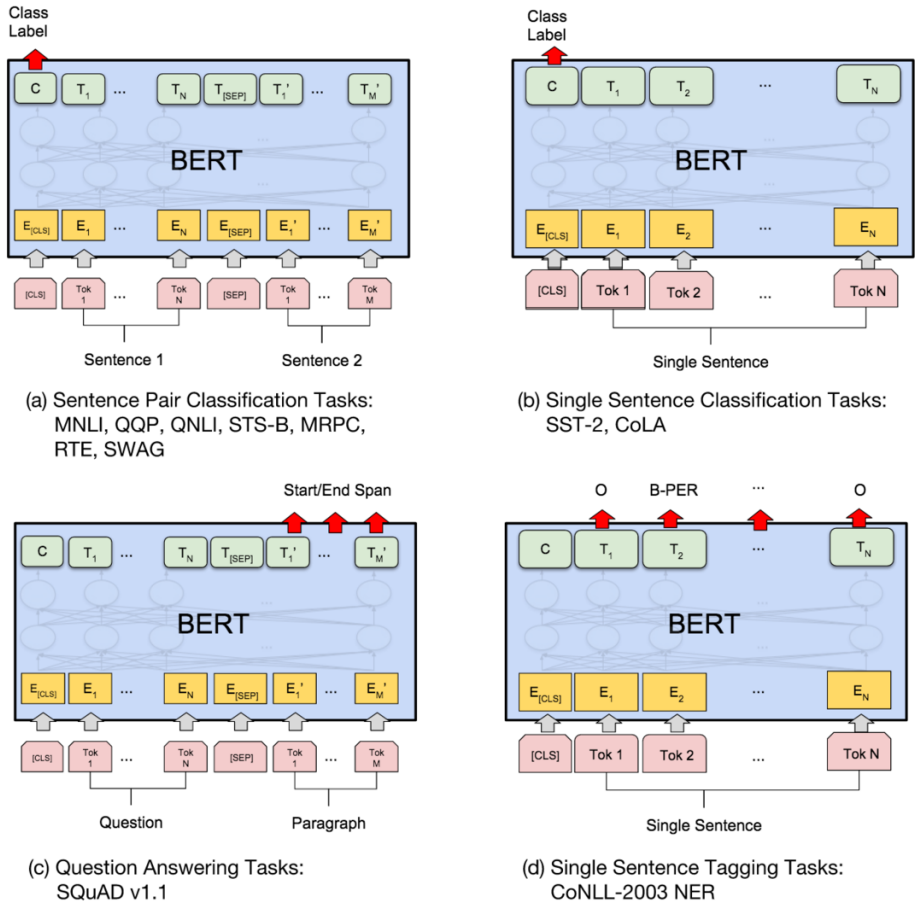
より高度なコンテキストを扱うことができる言語モデルが必要

課題と今後のテーマ

1. 表現の統一　→

＜今後のテーマ＞

BERT：前後の文脈・文法上の係りを理解する



・BERTはTransformerのEncoderを複数重ねたもの

・マスク化言語モデル：文中の一部の単語をマスクしてそれを予測する

・次文予測：２つ目の文が入力文の続く文であるかどうかを予測する

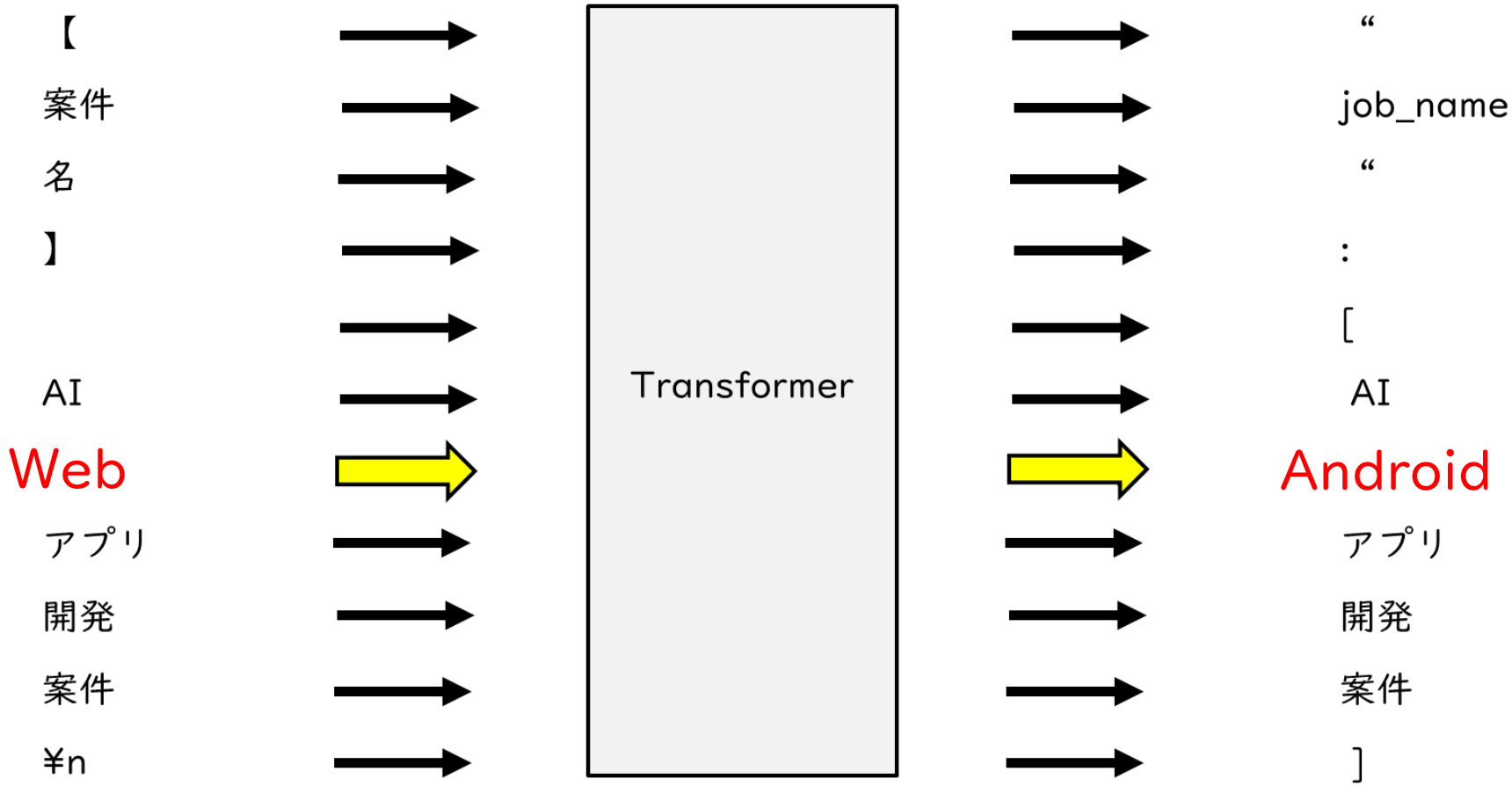
・ファインチューニングによって短時間で高精度の学習ができる

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina ToutanovaBERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805(2018)

課題

2. 翻訳の副作用

同一単語でも英語から日本語に翻訳するような作業をしているため、時々翻訳ミスが発生する



課題と今後のテーマ

2. 翻訳の副作用　→　＜今後のテーマ＞ **機械読解**：質問に答えるという形で特定個所を直接抽出する

Context：

メール文

【案件名】 AI Webアプリ開発案件
【内容】 チャットボットを用いたWebアプリの設計・開発に携わっていただきます
【勤務地】 茅場町 （東京メトロ東西線・日比谷線）
【期間】 10月～12月 ※延長の可能性あり
【必須スキル】 Python, C#を使用した開発経験 自然言語処理に関する知識
【推奨スキル】 javascript, django Webアプリ開発の経験
【単価】 50～60万前後
【年齢】 40歳はじめまでを希望
【人数】 1名
【面談回数】 2回（弊社同席）
【その他】 勤怠，コミュニケーションに問題のない方

← Question：

この案件で必要とされるスキルは何か？

↓

→ Answer：

Python, C#を使用した開発経験、自然言語処理に関する知識

ご清聴ありがとうございました