大規模言語モデルの語彙的関係知識推定における日英間の比較調査

東北大学　乾研究室

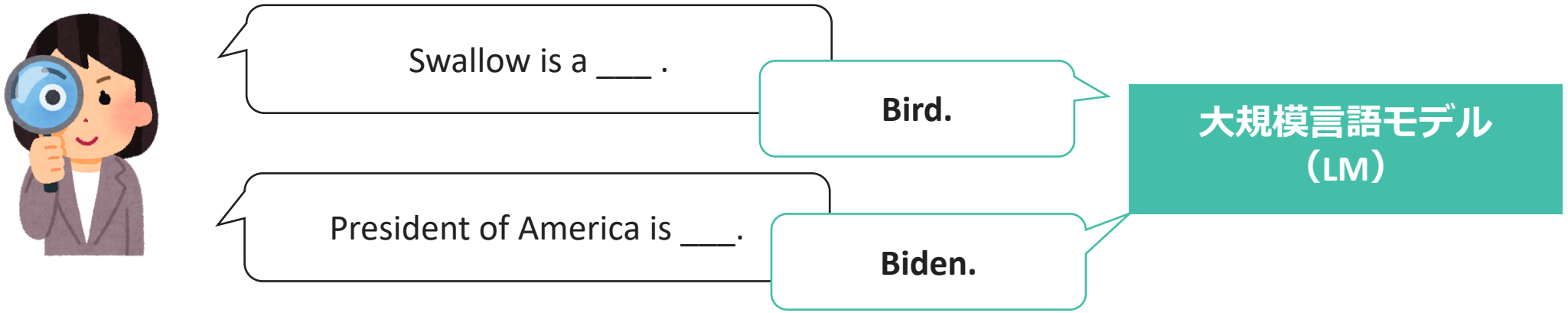
阿部 香央莉[1], 北山 晃太郎[1], 松田 耕史[2,1], 吉川 将司[1,2], 乾 健太郎[1,2]

E-mail: abe-k@tohoku.ac.jp

背景：大規模言語モデルの知識推定（Probing）

- BERTなどの大規模言語モデル（LM）の台頭は凄まじいが，
  LMの内部でどのような知識が蓄えられているかは未だ不明

→ 穴埋め形式の問題を利用した**LMの内部知識の推定**の研究が盛ん



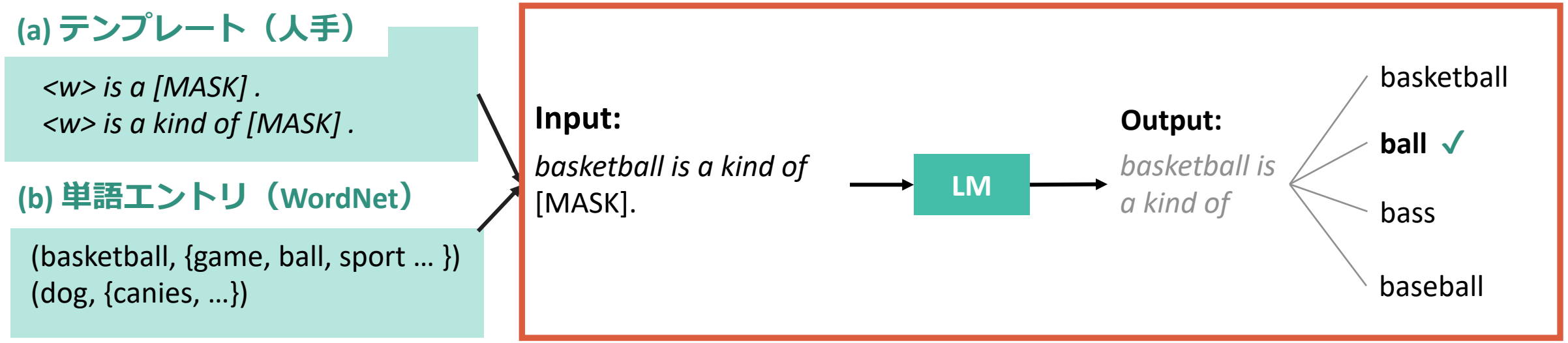
- たとえば，
  - LMは低頻度語に関する語彙知識を予測できない [Schick & Schütze, AAAI2020]
  - LMは否定（negation）に対して弱い [Ettinger+, ArXiv2019; Hosseini+, NAACL2021]

問題点：他言語における知識推定の検証が不十分

- 知識推定 (Probing) の研究の多くは，英語を対象とした議論にとどまっている
  - 多くの穴埋め形式データセットが提案されているが，ほとんどは英語
  - **日本語などの他言語を扱った研究は少ない**

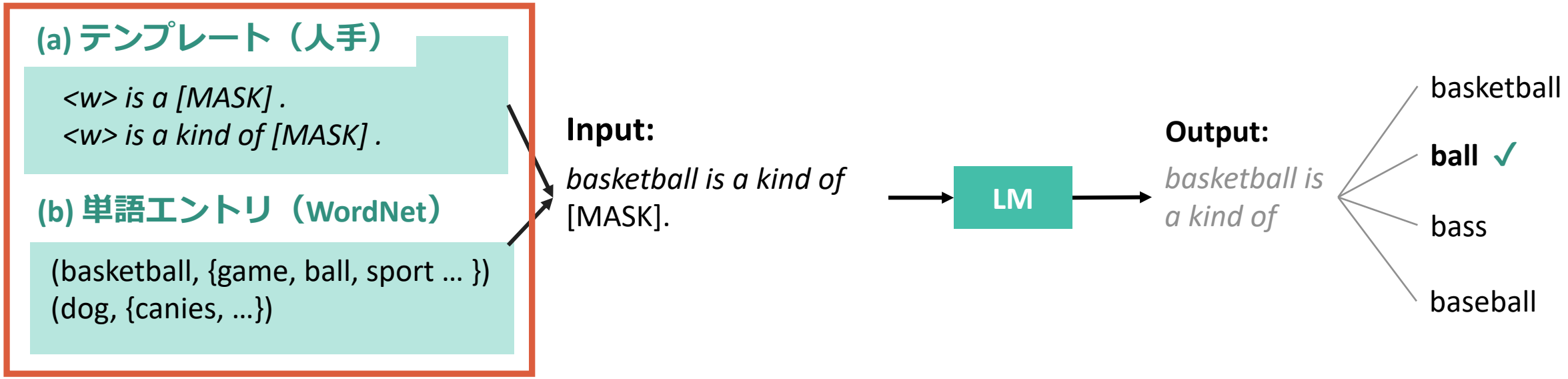
- **本研究の目的: 英語における語彙的関係知識推定の結果**[Schick & Schütze, AAAI2020]**と日本語の結果との比較**
  - 英語と言語類型が大きく離れた日本語においても，**頻度と予測性能の関係**が見られるか調査
  - **LMの知識推定のための日本語データセット**の自動構築
    - [Schick & Schütze, AAAI2020]らはデータ構築にWordNetを使用しており，他言語でも実験プロセスが再現しやすい

タスク：LMの語彙的関係知識推定



- Masked Language Modelタスクを解くように訓練されたLMには，文を穴埋めする能力がある
- LMの穴埋め能力を利用し，**対象となる知識を問うような穴埋め問題を作成**してその正否を調べる
  - 例：「basketball」は「ball」「game」「sport」の一種である（上位語）など

タスク：LMの語彙的関係知識推定



- 穴埋め問題のデータを作成する上で，以下の2つが必要
  (a) 穴埋め問題の雛形となる**テンプレート**（人手で作成）
  (b) (a)に当てはめる，キーワード（<w>）および正解となる単語群のペアで構成された**単語エントリ**（WordNetから抽出）

タスク：LMの語彙的関係知識推定 （英語 の場合）[Schick & Schütze, AAAI2020]

(a) テンプレート（人手）

<w> is a [MASK] .
<w> is a kind of [MASK] .

(b) 単語エントリ（WordNet）

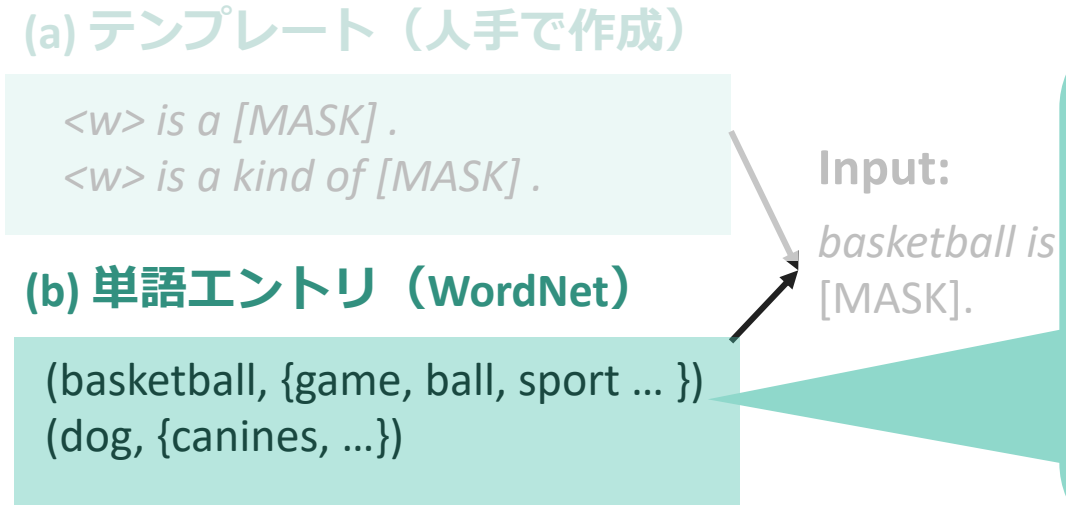
(basketball, {game, ball, sport … })
(dog, {canines, …})

- 各語彙関係（上位・対義・同位）に対応する，[MASK] トークンを含むテンプレートを作成

関係	テンプレート
上位語	*a <w> is a [MASK].* *“<w>” refers to a [MASK].* *a <w> is a kind of [MASK].*
対義語	*<w> is the opposite of [MASK].* *someone who is <w> is not [MASK].* *Something that is <w> is not [MASK].* *<w> is not [MASK].*
同位語	*<w> and [MASK].* *“<w>” and “[MASK]”.*

※同位語：同じ上位語を持つ単語同士の関係

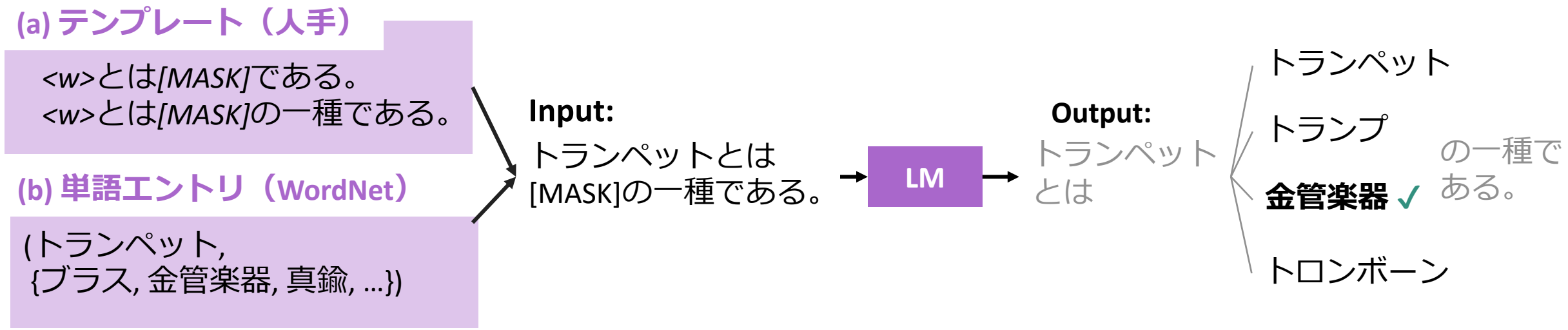
タスク：LMの語彙的関係知識推定 （英語 の場合）[Schick & Schütze, AAAI2020]



関係	キーワード	正解候補
上位語	basketball	game, ball, sport …
対義語	new	old
同位語	samosa	pizza, sandwich, salad

- WordNetから，各関係にある単語エントリを取得
  - このとき，正解候補となる単語はLMの分割器（tokenizer）をかけた後に1トークンで表されるもののみを使用
- キーワードの頻度で**低・中・高頻度（1~9, 10~99, 100>）**のサブセットに分割して実験

タスク：LMの語彙的関係知識推定（日本語 の場合）



- 英語の場合と同様に，以下の2つの日本語版を用意する
  - (a) 人手で作成した自然言語によるテンプレート
  - (b) WordNetから取得した単語エントリ

タスク：LMの語彙的関係知識推定 （日本語 の場合）

(a) テンプレート（人手）

<w>とは[MASK]である。
<w>とは[MASK]の一種である。

(b) 単語エントリ（WordNet）

(トランペット,
{ブラス, 金管楽器, 真鍮, …})

関係	テンプレート
上位語	<w>とは[MASK]である。 「<w> 」とは [MASK] の一種である。 「<w>」とは [MASK] のことを指す。
対義語	<w>と [MASK] は対の関係にある。 <w>と [MASK] は逆の関係にある。 <w>と [MASK] は反対の関係にある。 <w>の反対は [MASK] である。 <w>は [MASK] ではない。
同位語	<w>と [MASK]。 「<w> 」と「[MASK]」。 <w>と [MASK] の違い。 <w>と [MASK] の違いについて。

- 英語のテンプレートを参考に
  各関係に対応する
  日本語のテンプレートを作成

タスク：LMの語彙的関係知識推定 （日本語 の場合）

(a) テンプレート（人手）

<w>とは[MASK]である。
<w>とは[MASK]の一種である。

(b) 単語エントリ（WordNet）

(トランペット,
{ブラス, 金管楽器, 真鍮, …})

Input:
トランペ
[MASK]の

関係	キーワード	正解候補
上位語	トランペット	ブラス，金管楽器，真鍮，…
対義語	男	女，女性
同位語	ピンク	ブロンド，ブルー

- Open Multilingual WordNet [1]から英語と同様に単語エントリを取得
  - 上位語・同位語は，直接上位関係にあるWordNet synsetを使用
- 英語と同様に，キーワードの頻度に応じて**低・中・高頻度**のサブセットに分割

[1] http://compling.hss.ntu.edu.sg/omw/

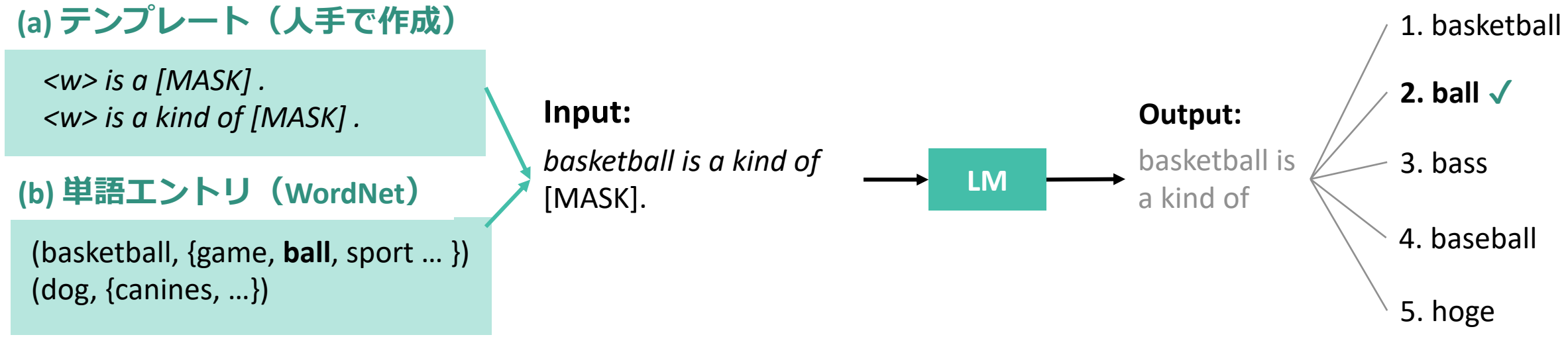
実験設定：データの統計量・使用した言語モデル

- WordNetから得られた単語エントリの総数

	英語 [Schick & Schütze, AAAI2020]			日本語		
	高頻度	中頻度	低頻度	高頻度	中頻度	低頻度
上位語	4,750	1,785	1,191	20,151	7,487	7,472
対義語	266	58	41	1,055	301	400
同位語	6,126	2,740	1,960	20,376	6,766	7,007

- 大規模言語モデル (LM)
  - 英語　：bert-base-uncased (https://huggingface.co/bert-base-uncased)
  - 日本語：bert-base-Japanese-v2 (https://huggingface.co/cl-tohoku/bert-base-japanese-v2)

評価指標: Mean Reciprocal Rank （逆順位平均）



- 各テンプレートに対するLMの予測を用いて，MRRを計算
  - たとえば，上記の例の場合，「*<w> is a kind of [MASK] .*」というテンプレートに対する**RR** (Reciprocal Rank, 逆順位) **= 2/5**
  - 最も良いRRとなるテンプレートを適用した際のRRの値を利用し，全ての単語エントリについて**RRの平均 (MRR)** を計算
- 予測数(top-k)は k=5とした

実験結果：日英での語彙的関係知識推定（MRR）

	英語			日本語		
	高頻度	中頻度	低頻度	高頻度	中頻度	低頻度
上位語	0.391	0.298	0.252	0.184	0.230	0.200
対義語	0.368	0.090	0.119	0.042	0.072	0.076
同位語	0.279	0.156	0.124	0.095	0.090	0.080

※英語の結果は予測数k=5で再実験した結果を記載

- 英語では，先行研究の通り，キーワードが低頻度になるほどMRRが下がっていく傾向が見られた
- 日本語では，**頻度とMRRの相関は見られなかった**

→ 言語横断的に調査を行うことの意義を示唆

分析：実際の上位語予測

- 「<w>は[MASK]の一種である。」というテンプレートでの結果

キーワード (<w>)	正解単語	BERTの予測単語（top-5）
フィッシュ・アンド・チップス	料理	寿司，以下，フィッシュ，**料理**，カクテル
ホッキョクグマ	クマ，熊	カニ，カエル，卵，魚，鳥類
排他的論理和ゲート	ゲート	**ゲート**，以下，これ，次，それ
レモンメレンゲパイ	パイ	**パイ**，以下，パン，菓子，ケーキ

- キーワードの分割がうまくできない例は予測に失敗しやすい
  - フィッ シュ・アンド・チップス → フィッシュ/・/アンド/・/ チップ/##ス
  - ホッキョクグマ → ホ/##ッキ/##ョ/##ク/グ/##マ
- 上位語をキーワード自体に含むエントリは予測しやすい → 予測が簡単な例？
  - ゲート，パイ

分析：実際の対義語予測

- 「<w>と[MASK]は対の関係にある。」というテンプレートでの結果

キーワード (<w>)	正解単語	BERTの予測単語（top-5）
男	女，女性	と，の，し，お，ん
右	レフト，左，左側，左手	と，し，の，ん ##ち
お子様	親	ママ，母，子供，子，女房
おとっつあん	マザー，実母，母，母親	と，で，プリン，は，ぴ

- **対義語に関して，日本語BERTではほぼ予測できない**
  - 特に意味を成さないひらがな一文字や，「反対」などの意味に相当する単語がよく出現する傾向が見られた
- なぜ予測できないのか？
  - 今のテンプレートではうまく予測できない（テンプレートの問題）？
    → AutoPrompt[Shin+, EMNLP2020] などの自動テンプレート生成方法の検討
  - 言語モデルの学習データ量不足？

分析：実際の同位語予測

- 『「<w>」と「[MASK]」。』というテンプレートでの結果

キーワード (<w>)	正解単語	BERTの予測単語（top-5）
ピンク	[ブロンド]，[ブルー，…]，[ブラウン，…]，[グリーン，…]	ピンク，赤，オレンジ，白，ブルー
犬	雌，[フォックス，狐，稲荷]，[ウルフ，オオカミ，狼]	猫，犬，人，ネコ，ウサギ
センチリットル	[l，l，L，ミル，cc，[UNK]]	リットル，センチ，センチメートル，キログラム，ミリ

- 正解と見なして良さそうな単語（「赤」「オレンジ」等）が予測できているにもかかわらず，WordNetによる正解が機能していない
  - 直接上位関係にあるsynsetのみに限定していることの弊害？
  - WordNet自体の質の問題？（not 一般的な表現が含まれている，一般的な表現が欠けている等）

⇒ データセットをより適切な評価ができる形に改善したい

分析：日本語における未知語トークンの出現

	日本語，低頻度	日本語，高頻度	英語，低頻度
未知語トークン出現数 / 全体	**198** / 7,472	23 / 20,151	0 / 4,750

※上位語での結果を示す

- 不正解となった例には，トークナイズを経てキーワードに未知語（[UNK]）トークンが含まれる例が見られた
  - 多くは，WordNetから抽出してきたエントリがBERTの文字語彙セット中に入っていない旧字体を含んでいる場合
  - 参考：bert-japaneseの文字vocabサイズは6,144
- LMの[MASK]トークンの予測においても[UNK]が出現する例も
  - 同位語予測の「センチリットル」の例など

まとめ・今後の展望

- まとめ
  - 英語による先行研究を基に，日本語においてLMの語彙的関係知識推定を行い，その結果を2言語間で比較した
  - キーワードの頻度と知識推定の精度の関係は見られず，言語横断的な調査の必要性が示唆された
  - 特に，日本語の低頻度語においては，LMにおける未知語（[UNK]）トークンの出現が多く見られ，これについて対処する必要がある

- 今後の展望
  - 現状のデータを日本語語彙知識推定用データセットとして妥当なものとするため，クラウドソーシング等によってデータを綺麗にする
  - 穴埋め形式による知識推定の方法論について，さらに深く検討する

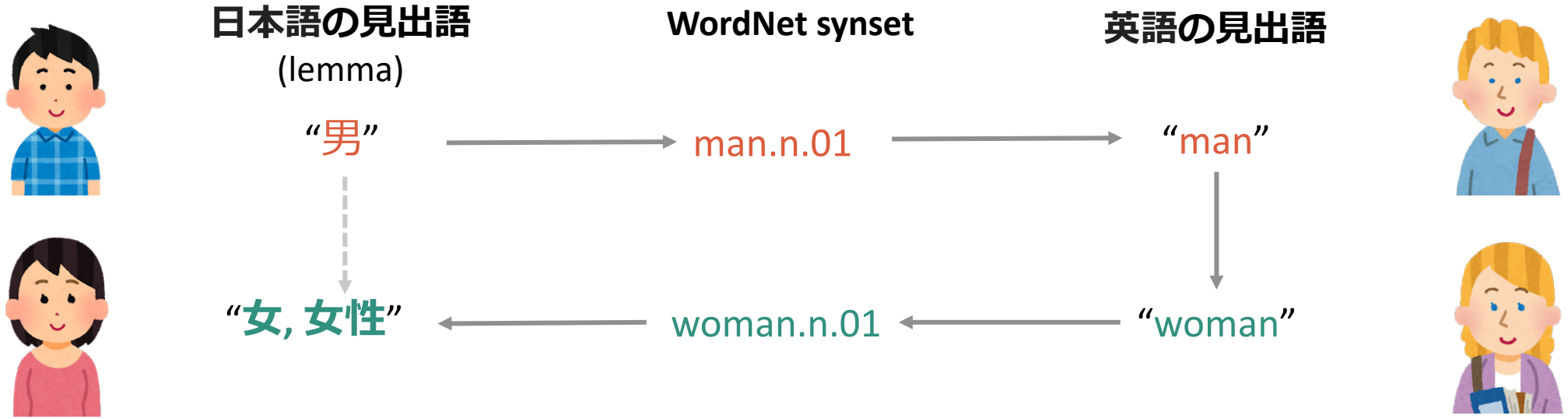
Appendix

日本語テンプレートの構築

英語	日本語
“<w>” refers to a [MASK].	「*<w>*」は*[MASK]*のことを指す。
“<w>” and “[MASK]”.	「*<w>*」と「*[MASK]*」。

- **テンプレートの要件**：対象となる単語をLMが予測できるような文
  - しかし，テンプレートが自然言語によるものである以上，対象の正解単語が必ず出力されることを担保しているわけではない
- 先行研究では，予測を助けるような小さな工夫がなされている
  - <w>や[MASK]の直前に冠詞を付加
  - <w>や[MASK]にダブルクォーテーションを付加
- 事実，本研究でも『「*<w>*」と「*[MASK]*」。』のようなテンプレートの方がうまく予測できる傾向にあった

注意点：日本語設定における対義語の取得



- 問題点：Open Multilingual WordNet において，
  日本語の見出語に対義語が整備されていない

→ **英語を経由して**対義語の日本語見出語を取得

- 完全に対応する対義語を取得するのは難しいと判断し，
  日本語の対義語エントリにおいては複数の正解を許容

注意点：日本語設定における対義語の取得

- 問題点：Open Multilingual WordNet において，
  日本語synsetに対義語が整備されていない
  - 2021年7月時点，他の言語（ポルトガル語，ドイツ語など）においても整備されていないことを確認済

- 一般的に対義語はsynsetではなく見出語 (lemma) に定義されている
  - 男 ⇔ 女，男性 ⇔ 女性などのニュアンスの違いを捉えるため

- 複数のエントリを許容することによる懸念・考察
  - 上記のニュアンスの違いを厳密に捉えられないため難易度が下がる可能性あり（が，現状そもそも対義語はほぼできていないためそれ以前の問題）
  - 対義の概念を意味的に理解していれば良いとするなら，これでも良いかも

分析：上位語予測における分析（正解単語の出力順位）

	正解単語の出力順位					
	1	**2**	**3**	**4**	**5**	**x**
全体	13.3	5.6	3.5	2.9	2.0	**72.6**
正解単語 in キーワード	**60.1**	14.5	5.1	3.4	2.7	14.3
Not 正解単語 in キーワード	4.4	4.0	3.2	2.8	1.9	**83.7**

- 「排他的論理和**ゲート**」「レモンメレンゲ**パイ**」のような，正解単語をキーワード中に含むものと含まないものに分けて各エントリの正解率を調査
- 正解単語をキーワード中に含むエントリは，容易に正解を予測できる
  - 反対に，含まない場合は顕著に予測できない

分析： 上位語予測における分析

	各キーワードのサブワード分割数					
	1	**2**	**3**	**4**	**5**	**6>**
高頻度	**19.4**	30.5	31.5	28.7	**51.7**	20.0
中頻度	2.1	29.5	34.0	32.0	31.7	33.3
低頻度	2.2	27.5	28.7	25.7	25.4	22.4

（表中の値は 各分割数 x 頻度において，1つでも正解単語を予測できたエントリの割合）

- 高頻度語は比較的どの分割数においても 正解率（%）が高い
- MRRの結果と照らし合わせると，日本語 BERT は正解単語の予測には成功しているものの，高順位で出力できていない

英語・日本語におけるキーワードの分割数

・**英語**エントリのサブワード分割数

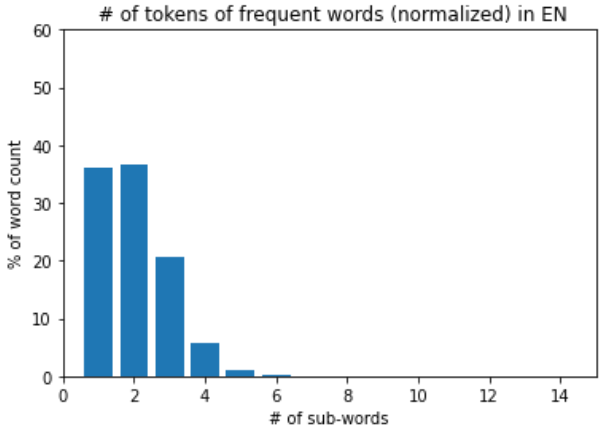
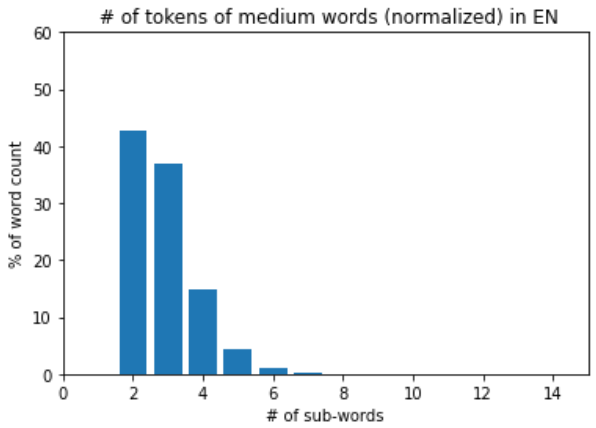
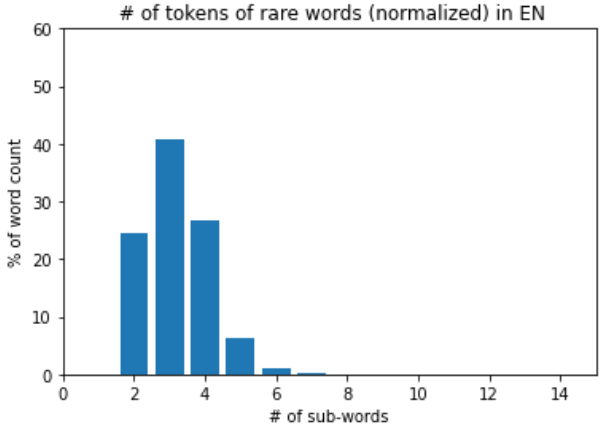
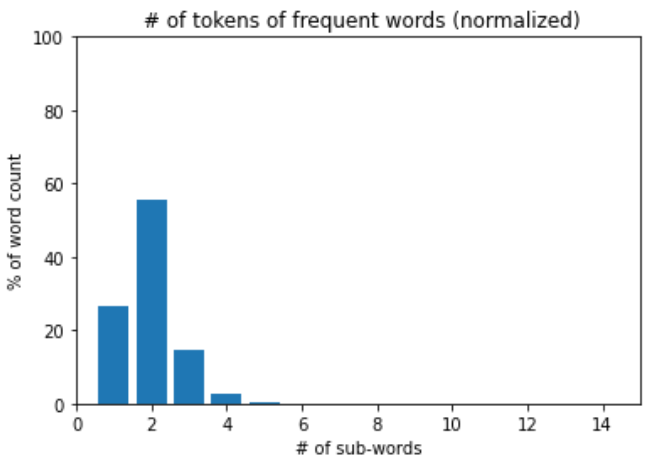
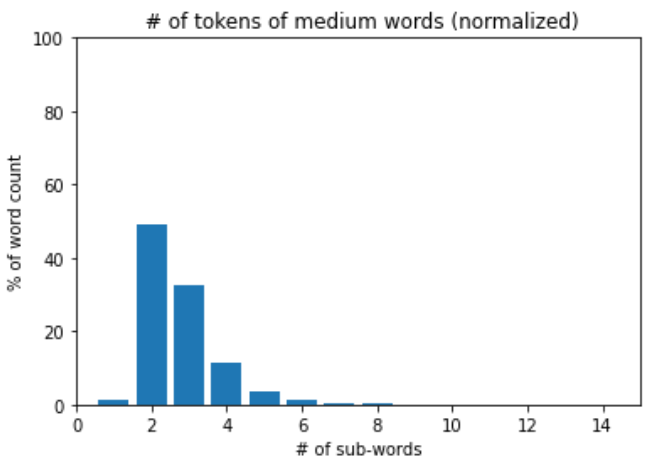
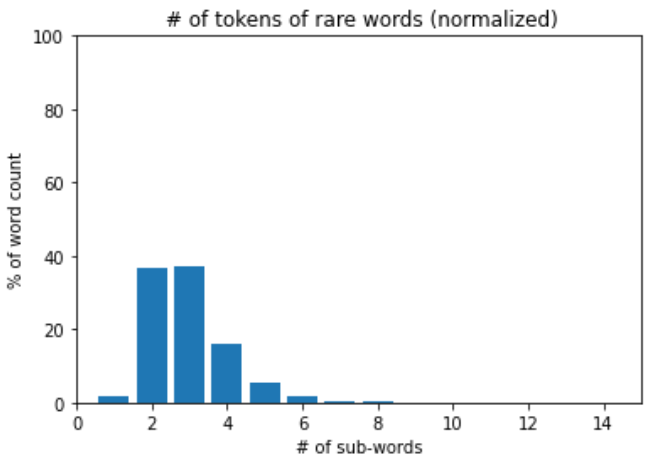
	1	2~4	5~
高頻度	3,145	5,490	109
中頻度	0	3,499	200
低頻度	0	5,243	444

・**日本語**エントリのサブワード分割数

	1	2~4	5~
高頻度	9,405	25,977	267
中頻度	176	13,884	824
低頻度	264	12,508	1,196

- 英語エントリでは中頻度以上でサブワード分割数1になるものはない
  日本語エントリでは一定数存在する
  - 中には難読漢字で[UNK]トークンになっているものも存在する

英語・日本語におけるキーワードの分割数

	高頻度	中頻度	低頻度
英			
日			

※縦軸は各頻度サブセットの総エントリ数で正規化したエントリ数の割合，横軸はサブワード分割数