

計算言語学₁₂

言語処理応用
(機械翻訳・多言語モデル)

東京大学生産技術研究所
吉永 直樹

site: <http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/class/cl/>

(選択)レポート課題について

- **課題 1:** 自然言語に関する主要国際会議論文 full paper (2016年以降, 8p以上)を読み内容をまとめる
<http://www.kamishima.net/archive/MLDMAImap.pdf> の大文字
 - NLP の国際会議 (上記参照): 3本～を 6p にまとめる or
 - NLP 以外の国際会議: 1本～を 3p にまとめる(推奨)
- **課題 2:** 自然言語に関する問題を新たに設定して取り組み結果を報告する (3～6p)
- 上記何れかを PDF で ynaga@iis.u-tokyo.ac.jp に送付
 - Subject: [Comp. Ling. Report 課題番号(1 or 2) 学籍番号]
 - 締切 2/6 23:59 JST 厳守 (遅延は例外なく不可)

補足: 課題 1について

- レポートには名前・学籍番号 + 以下を含めること
 - 論文タイトル、著者、会議名、出版年
 - 論文の内容のまとめ
(背景、目的、手法、実験結果、考察を簡潔に)
 - 論文の強い点を具体的に3つ (良く書いている、良く構成されているなどではだめ。なぜこの論文が採択されたのかを考え、その理由になりそうな点を挙げること。)
 - 論文の弱い点を具体的に3つ
- **重要:** こちらに読む論文を記入 (重複付加)
 - 早いもの勝ち
 - 指定本数以上の論文について被るのは OK (記入不要)

補足: 課題 2 について

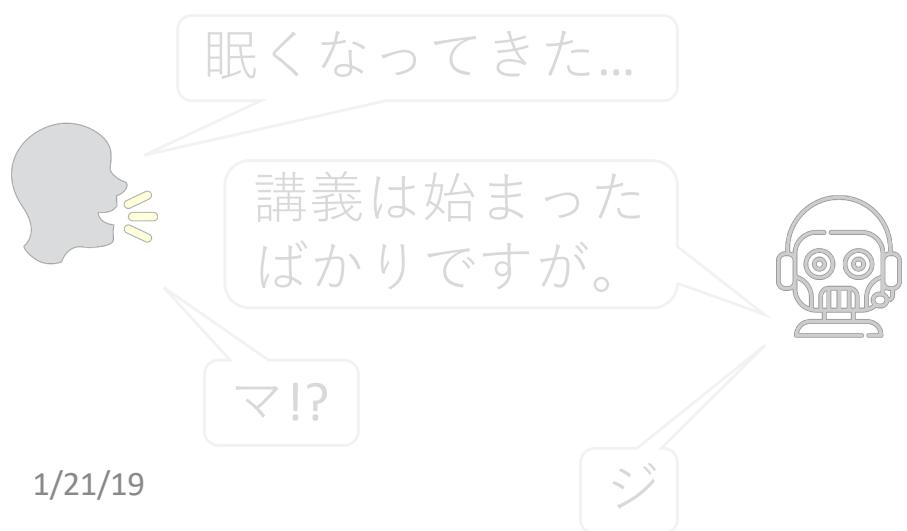
- レポートには名前・学籍番号 + 以下を含めること
 - 取り組んだタスク名
 - 取り組んだ内容のまとめ
(背景、目的、手法、実験結果、考察を簡潔に)
 - 実装したコードの場所
 - 評価では、タスクの新規性と意義を重視する
- **重要:** こちらで取り組むタスクを宣言 (重複禁止)
 - (彼らないとは思いますが) 早いもの勝ち

(エンドユーザを対象とした)言語処理応用

質問応答

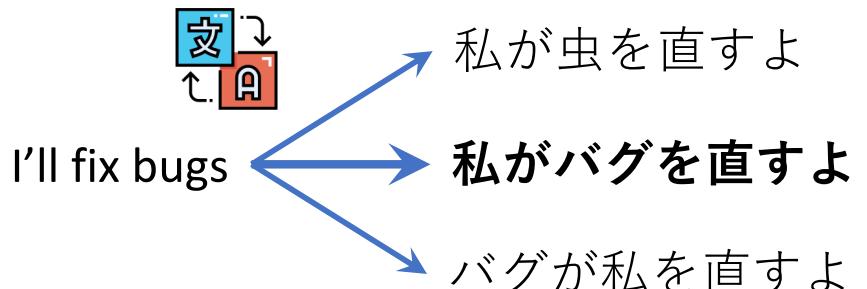


対話システム



1/21/19

機械翻訳



自動要約

John Forbes Nash Jr. (June 13, 1928 – May 23, 2015) was an American mathematician who made fundamental contributions to game theory, differential geometry, and the study of partial differential equations.^{[1][2]} Nash's work has provided insight into the factors that govern chance and decision-making inside complex systems found in everyday life. His theories are widely used in economics. Serving as a Senior Research Mathematician at Princeton University during the later part of his life, he shared the 1994 Nobel Memorial Prize in Economic Sciences with game theorists Reinhard Selten and John Harsanyi. In 2015, he also shared the Abel Prize with Louis Nirenberg for his work on nonlinear partial differential equations. John Nash is the only person to be awarded both the Nobel Memorial Prize in Economic Sciences and the Abel Prize. In 1959, Nash began showing clear signs of mental illness, and spent several years at psychiatric hospitals being treated for paranoid schizophrenia. After 1970, his condition slowly improved, allowing him to return to academic work by the mid-1980s.^[3] His struggles with his illness and his recovery became the basis for Sylvia Nasar's biography, *A Beautiful Mind*, as well as a film of the same name starring Russell Crowe as Nash.^{[4][5][6][7]} On May 23, 2015, Nash and his wife Alicia were killed in a car crash while riding in a taxi on the New Jersey Turnpike.

John Nash is a mathematical genius. 5

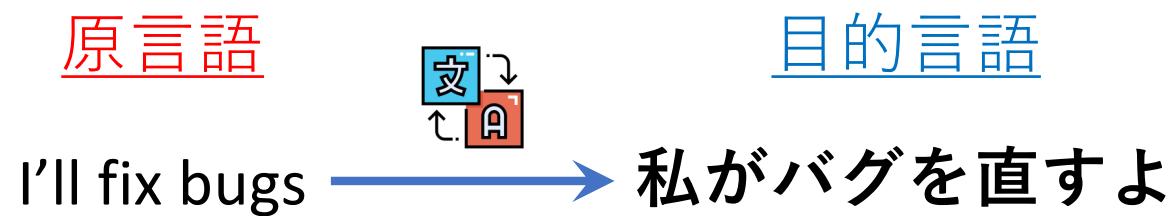
機械翻訳

Machine Translation

機械翻訳 (machine translation)

- ある言語(原言語)で書かれた文(章)を、同じ意味の他の言語(目的言語)の文(章)に変換

$$\hat{\boldsymbol{w}}^{(t)} = \underset{\boldsymbol{w}^{(t)}}{\operatorname{argmax}} \psi(\boldsymbol{w}^{(s)}, \boldsymbol{w}^{(t)})$$



- ・計算言語学に関する最古のタスク (1947-)
 - ・計算資源・言語資源の変化に伴い方法論もシフト
 - ・知識に基づく手法 → 用例ベース (1984-) → 統計ベース (1990-)

近年は深層学習のテストベットに (2014-)

機械翻訳の歴史: Warren Weaver の書簡 (1947)

Warren Weaver

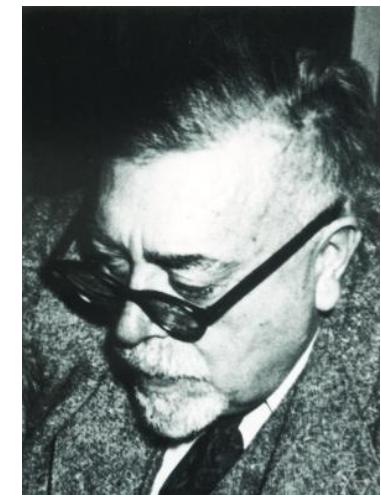


暗号解読のアナロジーとしての機械翻訳

... wonders if the problem of translation ... as a problem in cryptography. When I look at an article in Russian, I say "*This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.*"

... afraid the boundaries of words in different languages are too vague and the emotional and international connotations are too extensive to make any quasi mechanical translation scheme very hopeful.

Norbert Wiener

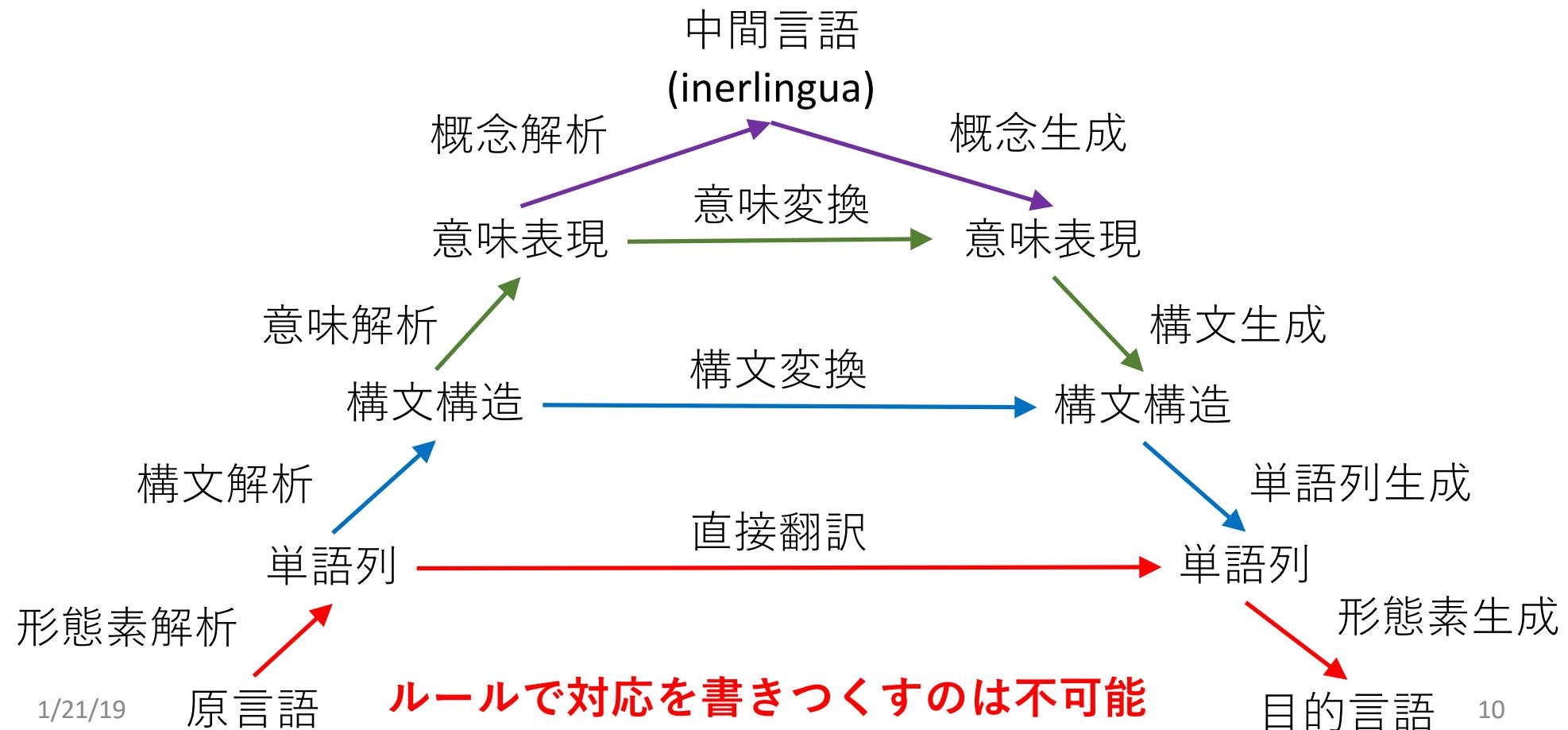


機械翻訳の難しさ：言語のズレ

- 語彙的なずれ
 - 多義語: bank (銀行, 土手), 月 (moon, month, Monday)
 - 粒度の違い: rice (米, ご飯, 稲), 飛ぶ (fly, jump)
- 構文的なずれ
 - 品詞: values are **positive** (正だ), **positive value** (正の)
 - 語順: **The professor was exhausted** (教授は疲れ果てた)
- 言語固有の問題
 - two dogs (二匹の犬), two birds (二羽の鳥)
 - 旅が好きだ (I like traveling)
 - 佐藤氏 (Mr./Ms. Satoh) は リンゴ (an/the apple(s)) を食べた

知識に基づく機械翻訳: Vauquois Pyramid

- 言語間のズレをどのレベルで吸収するか?
 - 抽象化されたレベルの方が対応関係の学習は容易だが、解析・生成のコストが発生（誤りが混入）



機械翻訳の訓練・評価データ

- 対訳コーパス (文単位の対訳データ)
<http://opus.nlpl.eu/>
注) 単語単位の対応は多くの場合ついていない
 - Hansards (英仏): カナダの国会議事録
 - EuroParl (21言語) [Kohén 2005]: ヨーロッパ議会の議事録
 - WIT3 (100k文対, 33言語): TED 講演 (字幕対訳)
 - OpenSubtitles: 映画 (字幕対訳)
 - NTCIR (3M文対; 日英): 特許文書
 - KFTT (443k文対, 日英): Wikipedia記事(京都関連)
 - ASPEC: 科学論文 (3M文対, 日英): 科学論文の抄録
- データに基づく機械翻訳に対訳コーパスは必須だが商用利用可能なものは極めて少ない

機械翻訳の用途

- 情報理解: 個人(あるいは計算機)が読む文書
 - 幅広い分野のテキストの翻訳が求められる
 - 荒い訳で良いが、**速い翻訳速度**が求められる
- 情報拡散: 公文書や広報、マニュアルの翻訳など
 - **高い翻訳精度と翻訳の一貫性(文体の統一)**が求められる
 - 翻訳速度は求められない
- 情報交換: 二者間の対話など
 - (比較的)高い翻訳精度と**リアルタイム性**が求められる

機械翻訳の評価 (1/2)

- 翻訳品質
 - **Fluency** (流暢さ): 翻訳文が言語として自然か
 - **Adequacy** (忠実さ): 翻訳文で原文の意味が保たれているか
- 例) 佐藤君は TensorFlow を操る

	Fluent?	Adequate?
To Sato it use TensorFlow	No	No
Sato debugs memory leaks	Yes	No
Sato uses TensorFlow	Yes	yes

近年は **参照訳を用いた自動評価** (BLEU [Papineni+ 2001]) が主流

- 翻訳速度: 一文辺りの実翻訳時間など

機械翻訳の評価 (2/2): 自動評価尺度

- BLEU [Papineni+ 2001]: コーパスレベルの自動評価
 - 参照訳との部分文字列の一致に基づく適合率を評価
 - Brevity Penalty (BP)により再現率を表現
 - 語順や同義性は評価されない(代名詞を間違うと致命的)

参照訳より短い訳を出す
場合のペナルティ

参照訳に含まれる
MT訳の*n*-gram数

$$\text{BLEU}(R, E) = \text{BP}(R, E) \cdot \frac{1}{N} \sum_{n=1}^N N \log p_n(R, E)$$

$$\text{BP}(R, E) = \min\left\{1, \frac{1 - \sum_i |\tilde{r}^{(i)}|}{1 - \sum_i |e^{(i)}|}\right\}$$

e に長さが最も近く,
かつ長さの短い参照訳

$$p_i(R, E) = \frac{\sum_j r_j^{(i)} \text{ と } e^{(i)} \text{ の共有 } n\text{-gram 数の最大値}}{\sum_i e_i \text{ 中の } n\text{-gram 数}}$$

データに基づく機械翻訳の方式

- 用例ベース機械翻訳 (Example-based MT; EBMT) [Nagao 1984]
 - 入力文と類似する用例を対訳コーパスから収集
 - 得られた用例を入力に合わせて修正, あるいは結合
- 統計的機械翻訳 (Statistical MT; SMT) [Brown+ 1993]
 - Noisy Channel Model [Brown+ 1990] に基づき, 入出力の部分構造(語/句/構文)の翻訳関係と出力の流暢さをモデル化
- ニューラル機械翻訳 (Neural MT; NMT) [Cho+ 2014]
 - 深層学習に基づく Encoder-Decoder モデルを用いて, 単語列間, あるいは原文間の写像を学習

統計的機械翻訳:

Noisy Channel Model (雑音のある通信路モデル)

- 基本的なアイデア: 流暢さと忠実さをスコア付け

$$\psi(\mathbf{w}^{(s)}, \mathbf{w}^{(t)}) = \psi_A(\mathbf{w}^{(s)}, \mathbf{w}^{(t)}) + \psi_F(\mathbf{w}^{(t)})$$

- Noisy Channel Model [Shannon 1949, Brown 1990]

- 入力からの出力の生成を事前確率と尤度でモデル化

$$\begin{aligned}\hat{\mathbf{w}}^{(t)} &= \operatorname{argmax}_{\mathbf{w}^{(t)}} p(\mathbf{w}^{(t)} | \mathbf{w}^{(s)}) \\ &= \operatorname{argmax}_{\mathbf{w}^{(t)}} \frac{p(\mathbf{w}^{(s)} | \mathbf{w}^{(t)}) p(\mathbf{w}^{(t)})}{p(\mathbf{w}^{(s)})}\end{aligned}$$

$$\operatorname{argmax}_{\mathbf{w}^{(t)}} p(\mathbf{w}^{(s)} | \mathbf{w}^{(t)}) p(\mathbf{w}^{(t)})$$

デコーダ 翻訳モデル 言語モデル

統計的機械翻訳: 単語アラインメント (1/3) IBM モデル1

- アイデア: 対訳文対から対訳単語対を推定

$$P(\mathbf{w}^{(s)} | \mathbf{w}^{(t)}) = \sum_{\mathbf{a}} p(\mathbf{w}^{(s)}, \mathbf{a} | \mathbf{w}^{(t)})$$

\mathbf{a} 原言語の各語が目的言語のどの語と対応するか

- IBM モデル1 [Brown+ 1993]による単語アラインメント

$$\hat{\mathbf{a}} = \operatorname{argmax}_{\mathbf{a}} p(\mathbf{w}^{(s)} | \mathbf{w}^{(t)})$$

目的言語から原言語への
一対多のアラインメントをモデル化

$$= \operatorname{argmax}_{\mathbf{a}} p_l(|\mathbf{w}^{(s)} = J | \mathbf{w}^{(t)}) \prod_{j=1}^J p_a(a_j | a_1^{j-1}, w_1^{j-1(s)}, \mathbf{w}^{(t)}) p_t(\mathbf{w}_j^{(s)} | a_1^j, w_1^{j-1(s)}, \mathbf{w}^{(t)})$$

$$= \operatorname{argmax}_{\mathbf{a}} \frac{\epsilon}{(|\mathbf{w}^{(t)}| + 1)^J} \prod_{j=1}^J p_t(\mathbf{w}_j^{(s)} | \mathbf{w}_{a_j}^{(t)})$$

ビタビアルゴリズムで
効率的にデコード可能

原言語中の語に対応する
訳語の添字は一様分布

原言語中の語は
訳語一つに対応

統計的機械翻訳: 単語アラインメント (2/3)

Hidden Markov alignment-Model (HMM)

- HMMに基づく単語アラインメント
 - アラインメント間の依存関係を考慮

$$P(\mathbf{w}^{(s)}, \mathbf{a} | \mathbf{w}^{(t)}) = P(\mathbf{w}^{(s)} = J | \mathbf{w}^{(t)} = I) \prod_{j=1}^J P(a_j | a_{j-1}, I) P(\mathbf{w}_j^{(s)} | \mathbf{w}_{a_j}^{(t)})$$

- アラインメントの確率は単語間の移動距離で算出

$$P(a_j = i | a_{j-1} = i', I) = \frac{C(i - i')}{\sum_{i''=1}^I C(i'' - i')}$$

統計的機械翻訳: 単語アライメント (3/3): パラメタの推定

- EM アルゴリズム $\hat{\theta} = \operatorname{argmax}_{\theta} \prod_{(\mathbf{w}^{(s)}, \mathbf{w}^{(t)}) \in \mathcal{D}} \sum_{\mathbf{a}} p_{\theta}(\mathbf{w}^{(s)}, \mathbf{a} | \mathbf{w}^{(t)})$
- E-step: 現パラメタ θ に基づき θ に関する期待値計算

<i>green</i>	<i>house</i>	<i>green</i>	<i>house</i>	<i>the</i>	<i>house</i>	<i>the</i>	<i>house</i>
1/3	1/3	1/3	1/3	1/3	1/3	1/3	1/3
<i>casa</i>	<i>verde</i>	<i>casa</i>	<i>verde</i>	<i>la</i>	<i>casa</i>	<i>la</i>	<i>casa</i>
$p_{\hat{\theta}_0}(\mathbf{a}, \mathbf{w}^{(s)} \mathbf{w}^{(t)}) = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}$	$p_{\hat{\theta}_0}(\mathbf{a}, \mathbf{w}^{(s)} \mathbf{w}^{(t)}) = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}$	$p_{\hat{\theta}_0}(\mathbf{a}, \mathbf{w}^{(s)} \mathbf{w}^{(t)}) = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}$	$p_{\hat{\theta}_0}(\mathbf{a}, \mathbf{w}^{(s)} \mathbf{w}^{(t)}) = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}$	$p_{\hat{\theta}_0}(\mathbf{a}, \mathbf{w}^{(s)} \mathbf{w}^{(t)}) = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}$	$p_{\hat{\theta}_0}(\mathbf{a}, \mathbf{w}^{(s)} \mathbf{w}^{(t)}) = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}$	$p_{\hat{\theta}_0}(\mathbf{a}, \mathbf{w}^{(s)} \mathbf{w}^{(t)}) = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}$	$p_{\hat{\theta}_0}(\mathbf{a}, \mathbf{w}^{(s)} \mathbf{w}^{(t)}) = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}$
$p_{\hat{\theta}_0}(\mathbf{a} \mathbf{w}^{(s)}, \mathbf{w}^{(t)}) = \frac{1/9}{2/9} = \frac{1}{2}$							
$C(\text{casa} \text{green}) = 1/2$	$C(\text{verde} \text{green}) = 1/2$	$C(\text{la} \text{green}) = 0$	$C(\text{green}) = 1$	$C(\text{casa} \text{house}) = 1/2 + 1/2$	$C(\text{verde} \text{house}) = 1/2$	$C(\text{la} \text{house}) = 1/2$	$C(\text{house}) = 2$
$C(\text{case} \text{the}) = 1/2$	$C(\text{verde} \text{the}) = 0$	$C(\text{la} \text{the}) = 1/2$	$C(\text{the}) = 1$				

- M-step: 期待値に基づきパラメタ θ を最尤推定

$p_{\theta}(\text{casa} \text{green}) = 1/2$	$p_{\theta}(\text{verde} \text{green}) = 1/2$	$p_{\theta}(\text{la} \text{green}) = 0$
$p_{\theta}(\text{casa} \text{house}) = 1/2$	$p_{\theta}(\text{verde} \text{house}) = 1/4$	$p_{\theta}(\text{la} \text{house}) = 1/4$
$p_{\theta}(\text{case} \text{the}) = 1/2$	$p_{\theta}(\text{verde} \text{the}) = 0$	$p_{\theta}(\text{la} \text{the}) = 1/2$

E-step → M-step
を繰り返す

統計的機械学習: 句に基づく翻訳モデル 翻訳モデルの推定

- 多対多の句アラインメントをモデル化

$$P(\mathbf{w}^{(s)} | \mathbf{w}^{(t)}) = \sum_{\mathbf{a}} p(\mathbf{w}^{(s)}, \mathbf{a} | \mathbf{w}^{(t)})$$

真面目に計算するの
は計算量的に困難

\mathbf{a} 原言語の句が目的言語のどの句と対応するか

- 単語アラインメントの対称化により効率的に計算
 - Intersection を求め union のアラインメントを追加 [Ozh 2003]

	Maria	no	dio	una	bofetada	a	la	bruja	verde	
Mary										
did										
not										
slap										
the										
green										
witch										

	Maria	no	dio	una	bofetada	a	la	bruja	verde	
Mary										
did										
not										
slap										
the										
green										
witch										

	Maria	no	dio	una	bofetada	a	la	bruja	verde	
Mary										
did										
not										
slap										
the										
green										
witch										

統計的機械学習: 構文に基づく翻訳モデル

- 同期文法 (synchronous grammar) [Chang 2005, Nesson+ 2007] を用いた機械翻訳

- 原言語の構文解析 = 目的言語での生成

スペイン語 英語

$NP \rightarrow (DET_1 NN_2 JJ_3, DET_1 JJ_3 NN_2)$

$JJ \rightarrow (enojado_1, angry_1)$

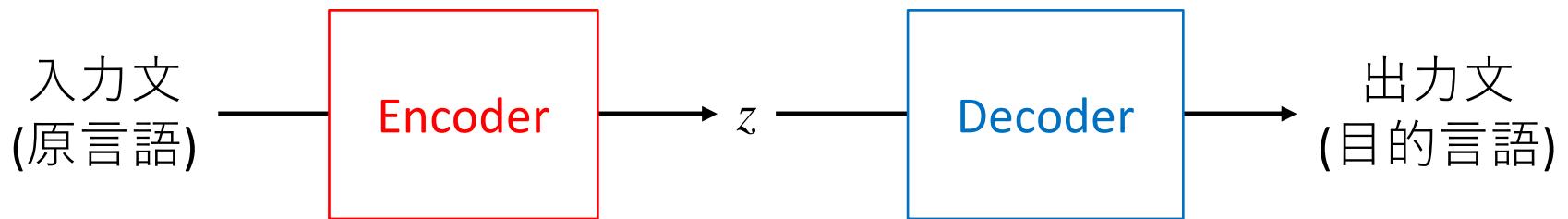
- 文法学習:

- 単語のアラインメントを付与した対訳コーパスから学習

- Decoding: CKY アルゴリズム

ニューラル機械翻訳 (Neural Machine Translation; NMT) Encoder Decoder モデル

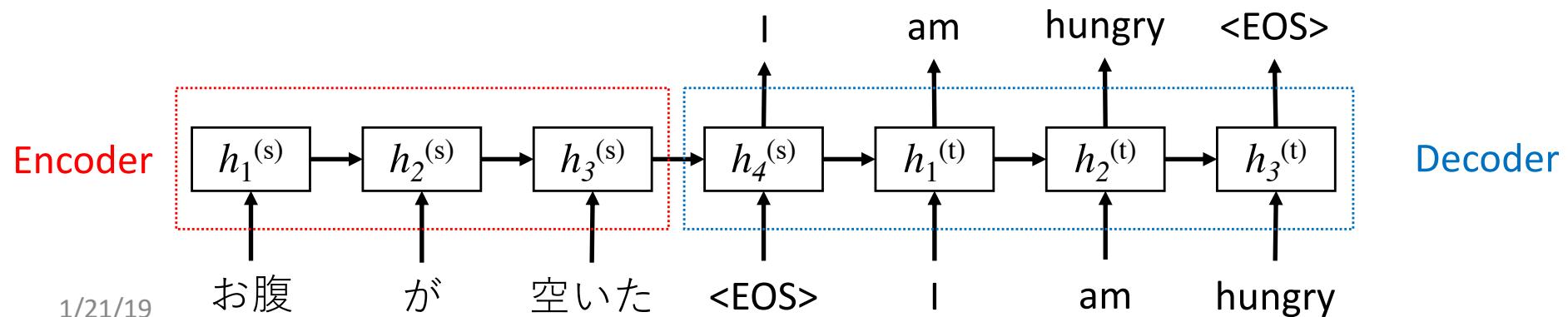
- **Encoder-Decoder モデル** [Cho 2004]
 - **Encoder**: 原言語の文を実数値ベクトル(の集合) z に変換
 - **Decoder**: z から目的言語の文を生成(符号化)



- Encoder, Decoder に用いる深層学習モデルにより、
様々な亜種が存在
 - **seq2seq** [Sutskever+ 2014, Luong+ 2015]: **RNN + 注意機構**
 - Transformer [Vaswani+ 2017]: 自己注意機構

ニューラル機械翻訳 (Neural Machine Translation; NMT) Sequence-to-Sequence モデル

- Sequence-to-Sequence (seq2seq) モデル [Sutskever+ 2014]
 - RNN (LSTM, GRU etc.) に基づく Encoder / Decoder を利用

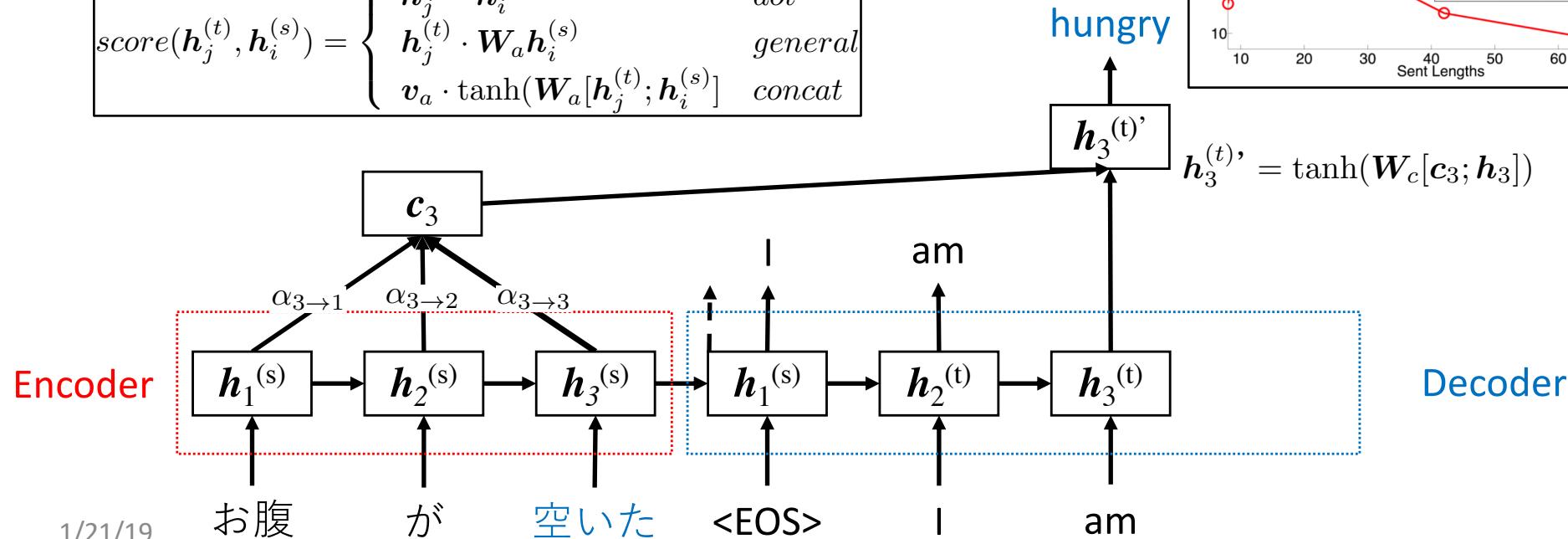


ニューラル機械翻訳 (Neural Machine Translation; NMT) Sequence-to-Sequence モデル + 注意機構

- Sequence-to-Sequence (seq2seq) モデル [Sutskever+ 2014]
 - RNN (LSTM, GRU etc.) に基づく Encoder / Decoder を利用
- 注意機構 [Luong+ 2015]: Encoder の情報を選択的に利用

$$\alpha_{j \rightarrow i} = \frac{\exp(score(\mathbf{h}_j^{(t)}, \mathbf{h}_i^{(s)}))}{\sum_i \exp(score(\mathbf{h}_j^{(t)}, \mathbf{h}_i^{(s)}))}$$

$$score(\mathbf{h}_j^{(t)}, \mathbf{h}_i^{(s)}) = \begin{cases} \mathbf{h}_j^{(t)} \cdot \mathbf{h}_i^{(s)} & dot \\ \mathbf{h}_j^{(t)} \cdot \mathbf{W}_a \mathbf{h}_i^{(s)} & general \\ \mathbf{v}_a \cdot \tanh(\mathbf{W}_a [\mathbf{h}_j^{(t)}; \mathbf{h}_i^{(s)}]) & concat \end{cases}$$



未知語への対応: サブワード

- デコーダは超多値分類を解く(softmax)
 - 学習・テスト時間の爆発
 - 語彙サイズの制限による未知語の増大(精度低下)
- 低頻度語をより短い部分文字列(サブワード)の系列で表現し、語彙サイズ・未知語を削減
 - 例) 足利義満=足利義+満, lower=low+er
- サブワード学習手法
 - Byte-pair encoding (BPE) [Gage 1994, Sennrich+ 2016]
 - ユニグラム言語モデル [Kudo 2018]

Decoding (復号化)

- 統計的機械翻訳の Decoding は非常に複雑 (省略)
- ニューラル機械翻訳の Decoding
 - Greedy search: $w_j^{(t)} = \operatorname{argmax}_{w \in V} p(w | \mathbf{w}^{(s)}, w_1^{j-1(t)})$
 - ビームサーチ: 各時刻で上位 b 個の解を保持
 - 深層学習の不安定性のため複数モデルの結果を参照するアンサンブルを用いることも

$$w_j^{(t)} = \operatorname{argmax}_{w \in V} \frac{1}{N} \sum_{n=1}^N p_n(w | \mathbf{w}^{(s)}, w_1^{j-1(t)})$$

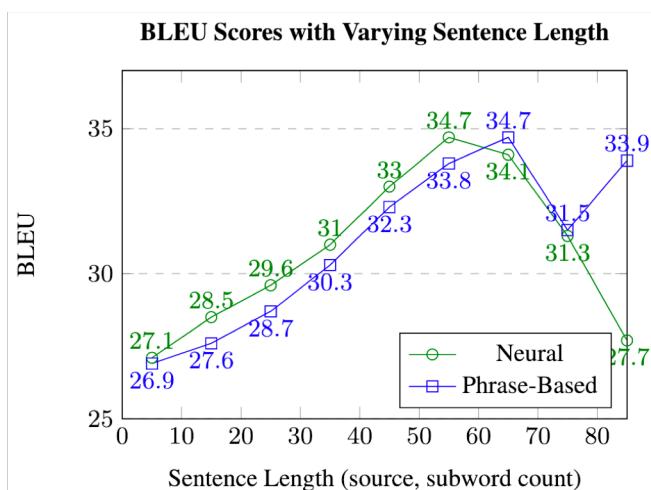
統計的機械翻訳 vs. ニューラル機械翻訳

[Koehn+ 2017]

NMT はドメイン依存性が強い

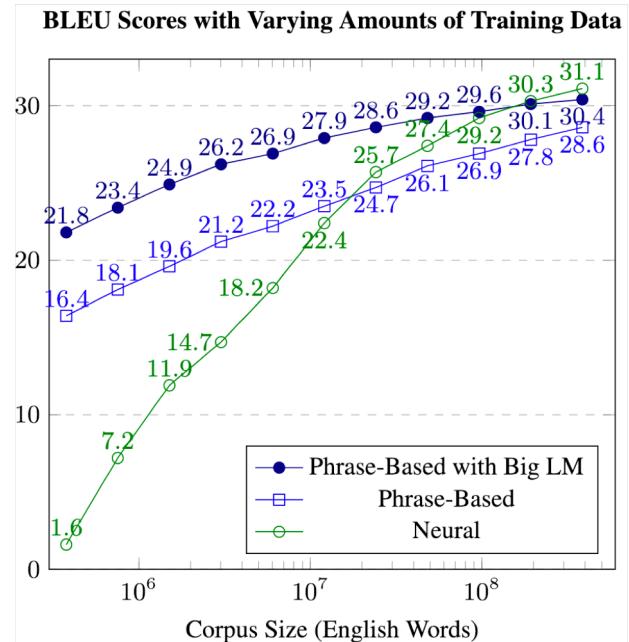
System ↓	Law	Medical	IT	Koran	Subtitles
All Data	30.5 32.8	45.1 42.2	35.3 44.7	17.9 17.9	26.4 20.8
Law	31.1 34.4	12.1 18.2	3.5 6.9	1.3 2.2	2.8 6.0
Medical	3.9 10.2	39.4 43.5	2.0 8.5	0.6 2.0	1.4 5.8
IT	1.9 3.7	6.5 5.3	42.1 39.8	1.8 1.6	3.9 4.7
Koran	0.4 1.8	0.0 2.1	0.0 2.3	15.9 18.8	1.0 5.5
Subtitles	7.0 9.9	9.3 17.8	9.2 13.6	9.0 8.4	25.9 22.1

NMT は長文に弱い



1/21/19

SMT は少データでも動作



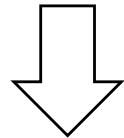
訳抜け、重複訳などNMT
特有の誤りへの対処

27

データに基づく機械翻訳とバイアス

- データに基づく生成はデータに存在するバイアスを増長

- Ms. Sato | Ito is a nurse. <-> 佐藤 | 飯尾さんは看護師だ
- Ms. Toda is a nurse. <-> 戸田さんは看護師だ
- Mr. Ueno is a doctor. <-> 上野さんは医者だ
- Mr. Sumi | Yata is a doctor. <-> 角 | 矢田さんは医者だ



- 坂井さんは看護師だ → Ms. Sakai is a nurse.
- 石井さんは医者だ → Mr. Ishii is a doctor

Mantra: マンガの超高性能な機械翻訳

<https://mntr.jp/>

- マンガ翻訳: マルチモーダルな機械翻訳
 - 異種フォントに頑健な文字認識技術
 - マンガ特有のトピックや文体に頑健な機械翻訳

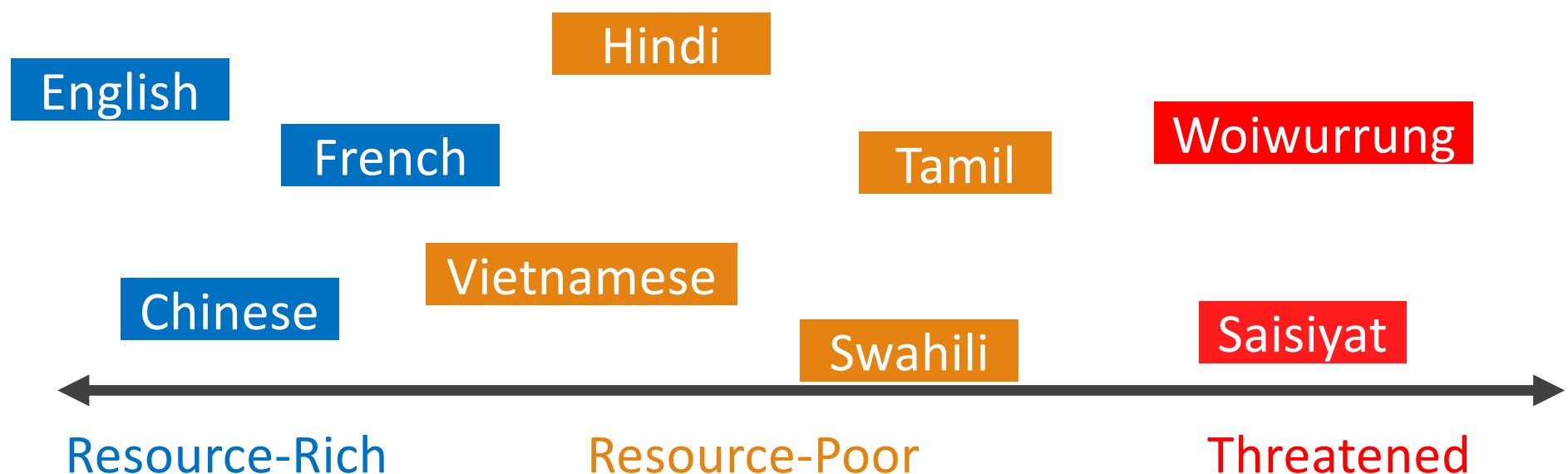


多言語モデル

Multilingual model

データに基づく言語処理の限界

- ほとんどの自然言語処理・計算言語学のタスクは機械学習で解かれる
 - 大規模な学習データを要求
 - 英語を除き、言語資源がない言語がほとんど



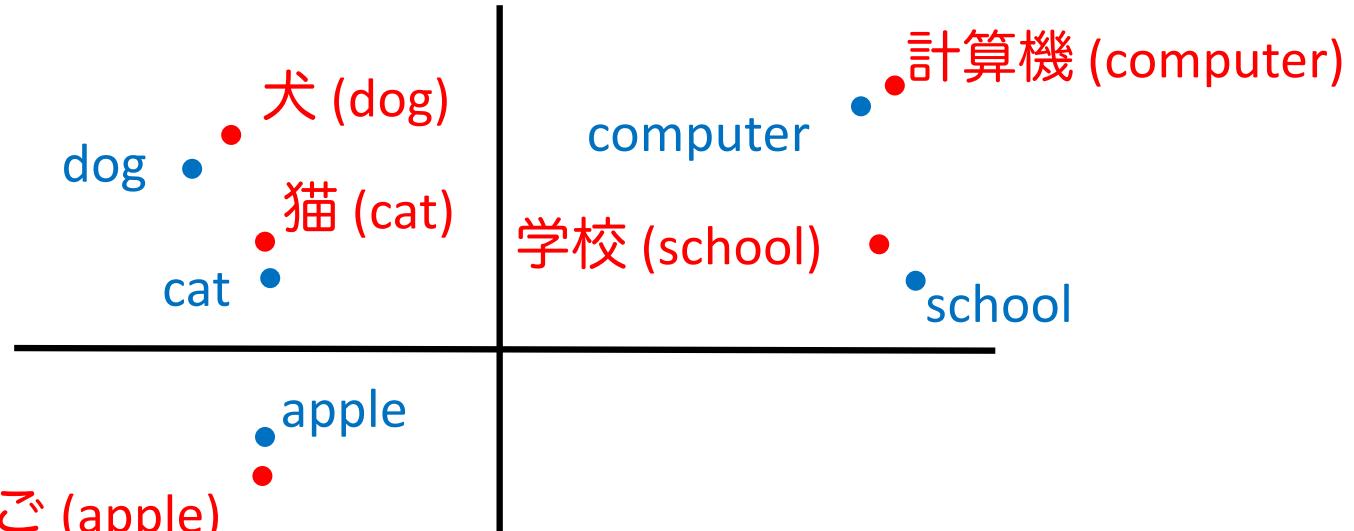
言語資源に乏しい言語における自然言語処理

- 言語資源の開発には原則、**その言語を母国語とする注釈者**が必要
 - 応用タスクの場合、クラウドソーシング対応可能
 - 基礎解析の場合、注釈付けに専門的知識が必要なため、研究者がいないとままならないケースも
- 多くのタスクで、統計的手法に基づく手法の精度は(学習データの少なさから)頭打ち・辞書等もない
 - 機械翻訳で翻訳しようにもまともな対訳コーパスもない

手法には暗黙の言語依存性がある場合が多く、近年は言語非依存の手法の開発が求められている
cf. Universal Dependencies: <https://universaldependencies.org/>

多言語モデル

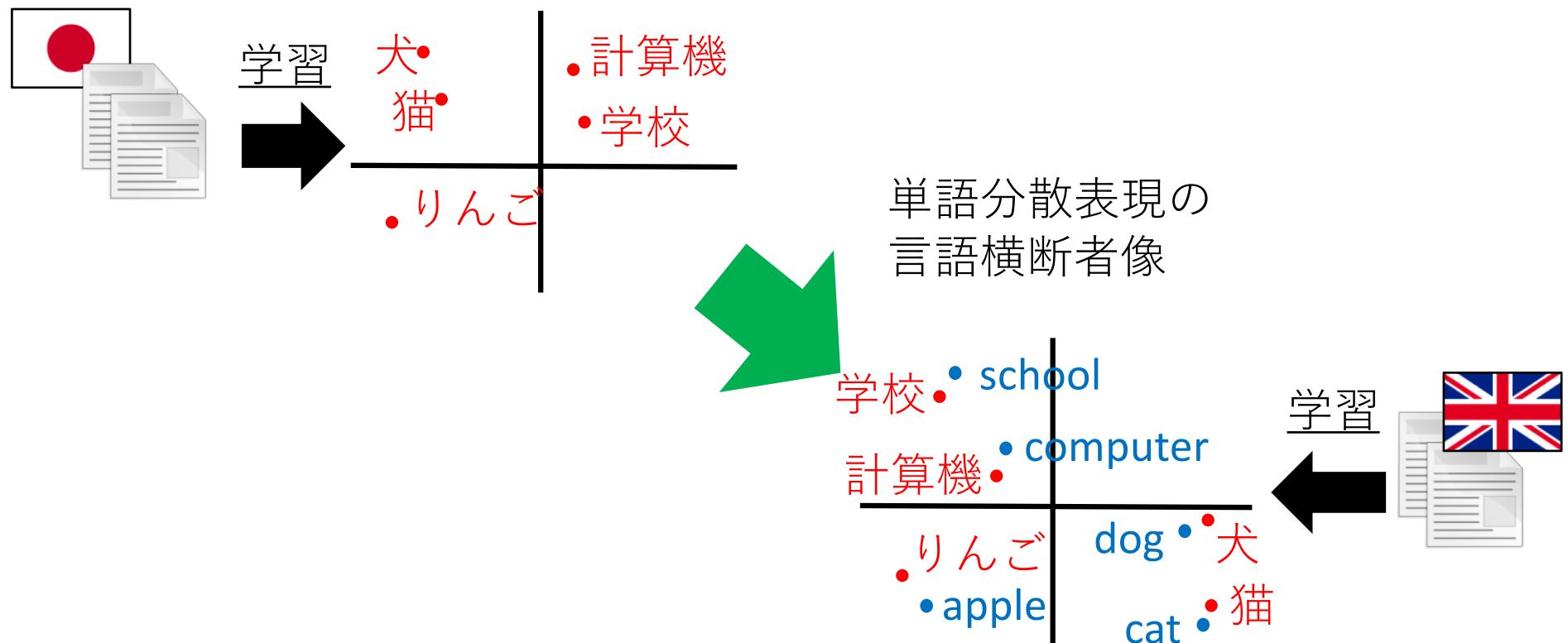
- 訓練した言語以外の言語にも適用できるモデルは実現可能か?
 - 語彙・語順・構文が異なるため困難
- 解決策: 多言語単語分散表現
 - 異なる言語を同じ意味空間に埋め込んだ単語の意味表現



深層学習モデルの単語埋め込み層を多言語単語分散表現に固定することで多言語モデルを獲得可能

多言語単語分散表現の獲得

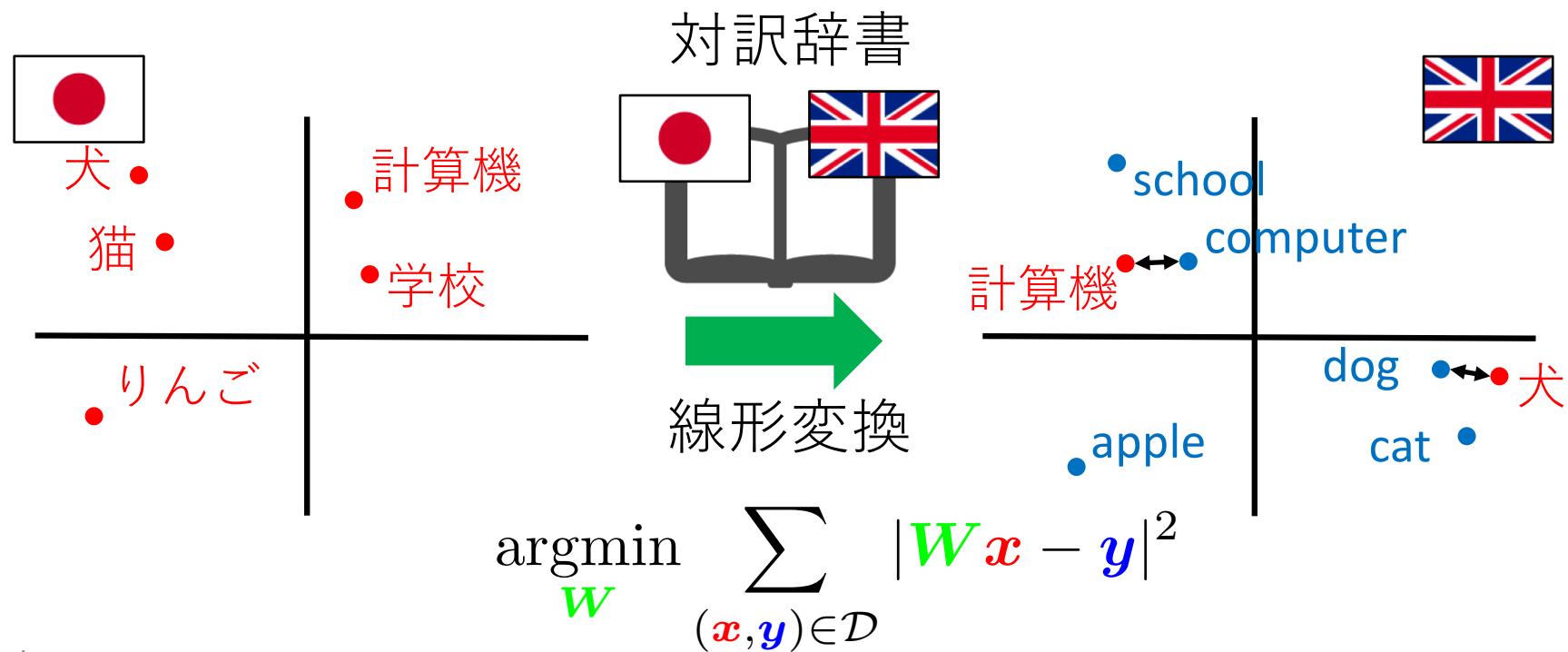
- 線形写像に基づく多言語単語分散表現の獲得
[Mikolov+ 2013, Xing+ 2015, Artexte 2018]



多言語単語分散表現の教師あり学習

[Mikolov+ 2013]

1. 単言語単語分散表現を skip-gram などで学習
2. 対訳辞書を教師データとして線形写像 W を学習
 - 変換を直行行列に制限すれば解析的に解ける [Xing+ 2015]
 - 分散表現の平均が $\mathbf{0}$ になるよう正規化し高精度化 [Arexte+ 2016]



多言語単語分散表現の教師なし学習

[Artexte 2018]

1. 単言語単語分散表現を skip-gram などで学習し,
正規化→平均を0化→正規化
2. 単語分散表現の空間が同型であることを利用し
て初期対訳辞書を構築（超低精度）
3. 自己学習により写像と対訳辞書を交互に最適化
 1. 写像を対訳辞書から学習
 2. 写像を単語分散表現に適用し近傍語から辞書を再学習

教師あり vs 教師なし多言語単語分散表現

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.	Mikolov et al. (2013)	34.93 [†]	35.00 [†]	25.91 [†]	27.73 [†]
	Faruqui and Dyer (2014)	38.40 [*]	37.13 [*]	27.60 [*]	26.80 [*]
	Shigeto et al. (2015)	41.53 [†]	43.07 [†]	31.04 [†]	33.73 [†]
	Dinu et al. (2015)	37.7	38.93 [*]	29.14 [*]	30.40 [*]
	Lazaridou et al. (2015)	40.2	-	-	-
	Xing et al. (2015)	36.87 [†]	41.27 [†]	28.23 [†]	31.20 [†]
	Zhang et al. (2016)	36.73 [†]	40.80 [†]	28.16 [†]	31.07 [†]
	Artetxe et al. (2016)	39.27	41.87 [*]	30.62 [*]	31.40 [*]
	Artetxe et al. (2017)	39.67	40.87	28.72	-
	Smith et al. (2017)	43.1	43.33 [†]	29.42 [†]	35.13 [†]
25 dict.	Artetxe et al. (2018a)	45.27	44.13	32.94	36.60
	Artetxe et al. (2017)	37.27	39.60	28.16	-
Init. heurist.	Smith et al. (2017), cognates	39.9	-	-	-
	Artetxe et al. (2017), num.	39.40	40.27	26.47	-
None	Zhang et al. (2017a), $\lambda = 1$	0.00 [*]	0.00 [*]	0.00 [*]	0.00 [*]
	Zhang et al. (2017a), $\lambda = 10$	0.00 [*]	0.00 [*]	0.01 [*]	0.01 [*]
	Conneau et al. (2018), code [‡]	45.15 [*]	46.83 [*]	0.38 [*]	35.38 [*]
	Conneau et al. (2018), paper [‡]	45.1	0.01 [*]	0.01 [*]	35.44 [*]
	Proposed method	48.13	48.19	32.63	37.33

まとめ

- 機械翻訳
 - 知識に基づく機械翻訳: Vauquois Pyramid
 - 用例に基づく機械翻訳: 情報検索的アプローチ
 - 統計的機械翻訳: Noisy Channel Model
 - 単語アラインメント (IBM モデル1)
 - フレーズベース機械翻訳
 - 構文に基づくニューラル言語モデル
 - ニューラル機械翻訳: Encoder Decoder モデル
 - 注意機構
- 多言語モデル
 - 多言語単語分散表現による語彙の吸収

本講義で扱えなかった話題

- 基礎解析:
 - 関係抽出, 含意関係認識, 文脈解析, 談話解析など
- 言語処理応用:
 - 自動要約, 誤り訂正, 作文自動採点
- マルチモーダル, 他分野との接近
 - Data-to-Text (キャプション生成), Text-to-Data
- 実状況に即した自然言語処理 (機械学習)
 - 教師なし学習, ノンパラメトリックベイズ,
ドメイン適用, 能動学習, クラウドソーシング