2013年7月30日
人工知能学会インタラクティブ情報アクセスと可視化マイニング研究会

ウェブマイニングを用いた英語学習支援

首都大学東京　小町守 <komachi@tmu.ac.jp>

永田昌明（NTT）, 荒瀬由紀（Microsoft Research Asia）
水本智也, 坂口慶祐, 澤井悠, 田尻俊宗, 林部祐太, 松本裕治（NAIST）

英語学習支援に関する共通タスクがたくさん開催されています

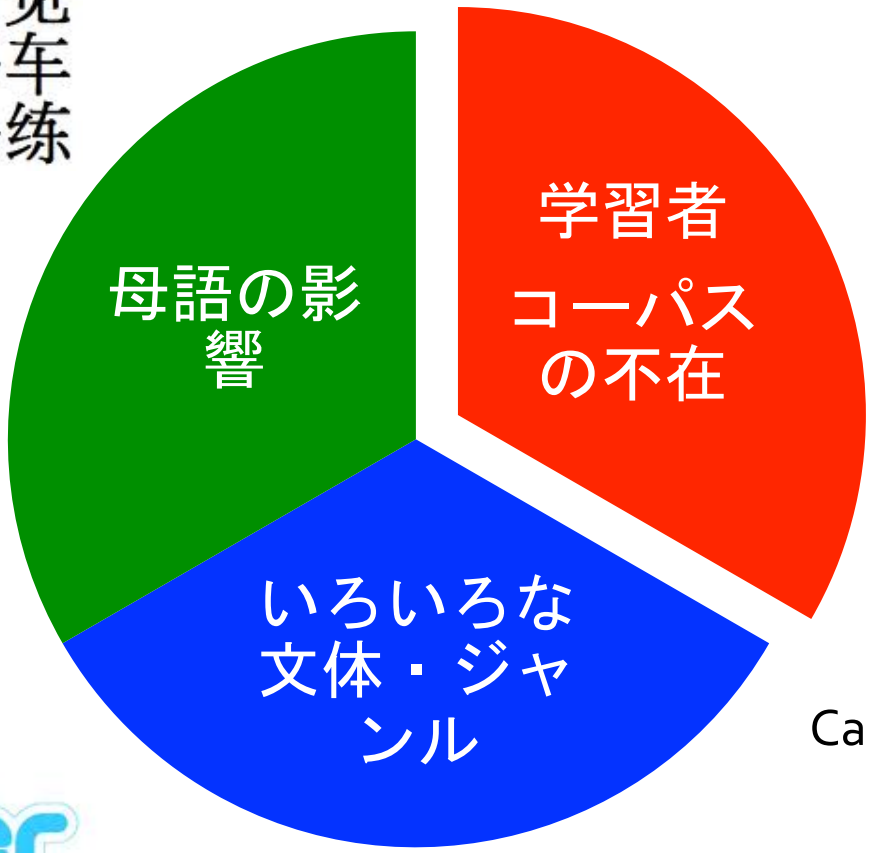
- Helping Our Own 2011 Pilot タスク
  - European Workshop on Natural Language Generation
  - 自然言語処理分野の論文の文法誤り訂正
  - ACL Anthology Reference Corpus
- HOO 2012
  - Building Educational Applications Workshop
  - 前置詞と限定詞（冠詞）の誤り訂正
  - Cambridge Learner Corpus First Certificate of English（ケンブリッジ英検）

今年も英語学習者支援に関する共通タスクが開催されました

- 母語推定共通タスク
  - Building Educational Applications Workshop
  - 英作文から英語学習者の母語を推定するタスク
  - TOEFL11 データセット（母語・トピックの均衡コーパス）
- 文法誤り訂正タスク
  - Conference on Natural Language Learning 2013
  - 限定詞、前置詞、数、動詞の形、主語動詞の一致
  - NUS Corpus of Learner English（シンガポール国立大、評価スクリプト同梱）

言語処理技術を作文支援に適用するにはいくつかの課題があります

簡→简　漢→汉
書→书　議→议
為→为　見→见
時→时　車→车
個→个　練→练



teressted|interested</#S> <#NT>|in</#NT>
<#W>Only <#NT>|on</#NT>
And <#NT>|on</#NT>
ome of them are in danger <#NT>|of</#NT>
<#NT>| On</#NT>
the phone, here is a list <#NT>|of</#NT>
<#NT> The|In t
chool wants to provide us <#NT>|with</#N
Please write <#NT>|to</#NT>
in order to let you know <#NT>|about</#
<#NT> First|At
regards</#S>|occurs</#RV> <#NT>|to</#NT>
orry</#TV>, even thinking <#NT>|about</#
o hours later, she learnt <#NT>|of</#NT>
were going to provide us <#NT>|with</#N
rry about</#RJ> this, but <#NT>|over</#N
it twice, 'Let's find out <#NT>|about</#
I am writing <#NT>|to</#NT>
theatre's restaurant, and <#NT>|to</#NT>
#RT> what they have asked <#NT>|for</#NT
RT> main European cities, <#NT>|for</#NT
I would like to suggest <#NT>|to</#NT>
In contrast <#NT>|to</#NT>
can <#RV>see|watch</#RV> <#NT>|for</#NT
FV>to ask|asking</#FV> me <#NT>|about</#
nclude|are included</#TV> <#NT>|in</#NT>
cerned, they didn't focus <#NT>|on</#NT>
ND> intimate relationship <#NT>|with</#N

Cambridge Learner Corpus

twitter

これまでいくつかのタスクで英語学習者支援に取り組んできました

- 識別学習を用いた英語学習者のための穴埋め問題生成 (ACL 2013)
- 学習者の誤り傾向を反映した英語動詞選択誤りへの訂正候補推薦 (ACL 2013)
- 英語学習者の作文の文法誤り訂正における学習者コーパスの分量の影響 (COLING 2012)
- 英語スペル訂正と品詞タグ付けの同時解析 (COLING 2012)
- 大域的文脈情報を用いた英語時制誤りの検出と訂正 (ACL 2012)
- 語学学習SNSの添削ログからの大規模語学学習者コーパスの抽出 (IJCNLP 2011)

語学学習 SNS の添削ログからの大規模英語学習者コーパスの抽出

水本智也, 小町守, 永田昌明, 松本裕治
IJCNLP 2011
言語処理学会年次大会 (2011)
人工知能学会論文誌 (2013)

田尻俊宗, 小町守, 松本裕治
ACL 2012
言語処理学会年次大会 (2012)

英語学習者の書いた作文には いろいろな誤りがあります

He works <prp crr="at">to</prp> <at crr="a">the</at> flowershop.

※
prp: 前置詞 preposition error
at: 冠詞 article error
crr: 訂正 correction

誤り検出: 誤りの場所を同定
誤り訂正: 誤り検出し、正しい表現に直す

KJ コーパス（KJ=甲南大学・教育測定研究所）

- 日本語学習者の170エッセイ（課題作文）
- 文法誤り情報と品詞・句情報を人手付与

人手による学習者コーパスがここ数年相次いでリリースされました

目的によっていろいろな情報が付与されている。学習者の母語・サイズなども様々。

- Cambridge Learner Corpus (CLC)
- International Corpus of Learner English (ICLE)
- NUS Corpus of Learner English (NUCLE)
- Konan-JIEM Learner Corpus (KJ Corpus)

<u>問題点</u>

- 公開されている分量が少ない（数千文）

着眼点: 大規模データがあるウェブから学習者コーパスを抽出します

Livemocha: 言語学習 SNS
決められたトピックについて作文すると、ネイティブスピーカーからコメントがもらえる

<u>問題点</u>

- コメントと本文が文対応していないので抽出しにくい

(Park and Levy, ACL 2011)
韓国人の TOEFL 練習作文をクロール

<u>問題点</u>

- 学習者の作文のみで、添削がされていない

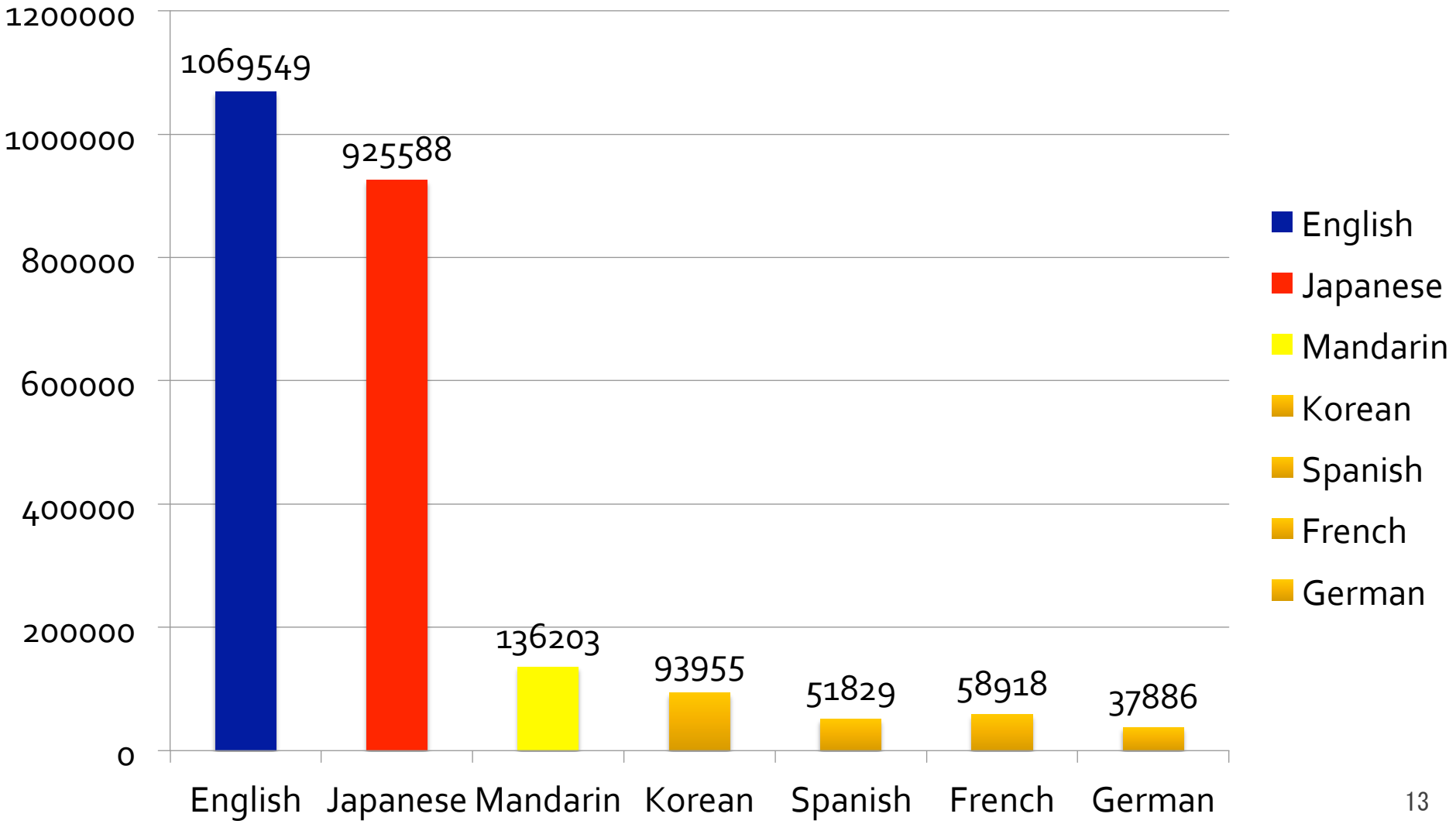
提案手法: Lang-8 の添削文対から言語学習者コーパスを構築します

Sentence written by a JSL learner	三人はそれぞれ自分の方式で感情を表れます。
Sentence corrected by an annotator1	三人はそれぞれ自分なりの表現で感情を表します。 Each of three expresses their feelings in their way of expression.
Sentence corrected by an annotator2	三人はそれぞれ自分なりに感情を表します。 Each of three expresses their feelings in their own way.

Lang-8: 言語交流 SNS http://lang-8.com/

- ユーザ数: 210,834 (2010年11月)
- ユーザが日記を投稿し、1文ごとにネイティブスピーカーが添削する。1つの文が複数の添削を受けることもある。

日英の添削文対は Lang-8 から大規模に抽出できます



Lang-8 英語学習者コーパス http://cl.naist.jp/nldata/lang-8

29,012ユーザによる100,051記事1,147,451文

<u>利点</u>

- 大規模に獲得可能
- 複数の添削候補を取得できる（こともある）

<u>欠点</u>

- 誤りの種類が分からない
- クラウドソースされた添削の品質は一定ではない (Han et al., LREC 2010)

Lang-8からの言語学習者コーパス抽出のまとめ

- 語学学習 SNS Lang-8 から100万文規模の英語学習者コーパスを抽出した。
- 品質は均一ではないが、ネイティブスピーカーによる添削がついている。
- 英語だけでなく、他の言語の学習者コーパスも抽出可能。
  - 日本語学習者コーパス (Mizumoto et al., IJCNLP 2011)
  - 日本語学習者ローマ字コーパス (Kasahara et al., WTIM 2011)

英語学習者の作文の文法誤り訂正における学習者コーパスの分量の影響

水本智也, 林部祐太, 小町守, 永田昌明, 松本裕治

COLING 2012 short
情報処理学会自然言語処理研究会（2012-11）

KJコーパスにおける誤り分布

- KJコーパスとCLCでは誤りの傾向が異なる

タイプ	%	タイプ	%
冠詞	19.23	名詞の語彙選択	7.04
名詞の単複	13.88	動詞の語彙選択	6.90
前置詞	13.56	代名詞	6.62
時制	8.77	動詞の一致	5.25

<u>疑問点</u>

1. 学習者コーパス vs ネイティブコーパス
2. 同じドメインの少量データ vs 違うドメインの大量データ
3. 誤り訂正についての母語の影響

統計的機械翻訳による学習者の作文の誤り訂正

- 学習者の作文から添削後の作文への変換を翻訳だと見なし、統計的機械翻訳によって誤り訂正する (Brockett et al., 2006)
  - 誤り訂正に関する特殊な知識は必要ない
  - 統計的機械翻訳のツールをそのまま適用可能
  - 統計的機械翻訳手法が改善されれば、それを援用することができる
  - 言語モデルの学習には大規模な生コーパス（例: Google N-gram）を用いることができる

関連研究: 統計的機械翻訳による学習者の作文の誤り訂正

- (Brockett et al., 2006)
  英語学習者の名詞の可算・不可算の誤りのみを対象としていた。また、学習者コーパスは人工データだった。
- (Mizumoto et al., 2011)
  Lang-8 コーパスを用いた統計的機械翻訳による誤り訂正手法だが、日本語学習者の作文が対象。また、誤りの種類別の結果を報告していない。

Lang-8 コーパスを用いた統計的機械翻訳による誤り訂正

主要な貢献

- 誤りの種類を限定しない初の大規模な実学習者コーパスを用いた英語学習者の誤り訂正

ツール

- アライメント: GIZA++ 1.0.5
- フレーズ抽出・デコーダ: Moses 2010-08-13

データ

- 訓練: KJ コーパス、Lang-8 コーパス
- 評価: KJ コーパス　（Precision, recall, F で評価）

少量のin-domainデータより out-domainでも大量にあると⬆

System	Training corpus	Recall	Precision	F-measure
Maximum entropy-based model	KJ Corpus	0.165	0.407	0.235
Phrase-based SMT	KJ Corpus	0.137	0.375	0.201
Phrase-based SMT	Lang-8 Corpus (390K)	**0.262**	**0.585**	**0.362**

Table 4: Result for preposition error correction on KJ Corpus.

MEモデルは NAACL 2012 ワークショップの HOO shared task で4位だったシステム（CLC FCE データセットで recall=18.44, Precision=34.88, F=24.12）で、表層、品詞、WordNet、統語情報および言語モデルを使用

※
フレーズベース SMT を KJ コーパスと Lang-8 コーパスで訓練したモデルは未実験。MEモデルは Lang-8 が大きすぎ訓練できず。

ほとんどの誤りで大規模データが有効

Training Corpus	KJ Corpus			Lang-8 Corpus (2K)			Lang-8 Corpus (390K)		
	Recall	Prec	F	Recall	Prec	F	Recall	Prec	F
article	0.187	0.531	0.277	0.187	0.571	0.282	**0.359**	**0.761**	**0.488**
noun number	0.207	0.603	0.308	0.136	0.671	0.226	0.199	**0.710**	0.311
preposition	0.137	0.375	0.201	0.092	0.319	0.143	**0.262**	**0.585**	**0.361**
tense	0.102	0.170	0.128	0.043	0.088	0.058	0.080	0.149	0.104
lexical choice of noun	0.035	0.114	0.054	0.033	0.152	0.054	**0.182**	**0.443**	**0.258**
lexical choice of verb	0.070	0.161	0.098	0.065	0.200	0.098	**0.192**	**0.324**	**0.241**
pronoun	0.075	0.220	0.112	0.040	0.143	0.063	0.150	**0.367**	**0.213**
agreement	0.236	**0.604**	0.340	0.125	0.483	0.199	0.228	0.469	0.307
adjective	0.151	0.326	0.206	0.056	0.286	0.094	**0.389**	**0.522**	**0.446**
verb other	0.089	0.139	0.109	0.147	0.333	0.204	**0.286**	**0.419**	**0.340**
adverb	0.265	0.450	0.333	0.214	0.429	0.286	0.292	0.432	0.349
conjunction	0.100	0.417	0.161	0.091	0.714	0.161	0.115	**0.546**	0.190
word order	0.500	0.025	0.048	0.667	0.050	0.093	**0.750**	0.075	0.136
noun other	0.182	0.222	0.200	0.143	0.167	0.154	**0.571**	**0.429**	**0.490**
auxiliary verb	0.056	0.167	0.083	0.100	0.400	0.160	0.100	**0.400**	0.160
other lexical choice	0.167	0.200	0.182	0.000	0.000	0.000	**0.357**	**0.455**	**0.400**
relative	0.111	0.250	0.154	0.182	0.667	0.286	0.091	**0.500**	0.154
interrogative	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Total	0.149	0.147	0.148	0.113	0.205	0.146	0.247	**0.275**	**0.260**

データ量が関係する誤り・しない誤り

Training Corpus	KJ	Lang-8						
		2K	10K	20K	100K	200K	300K	390K
article	0.277	0.282	*0.390	*0.420	*0.443	*0.459	*0.475	*0.488
noun number	0.308	0.226	0.214	0.238	0.270	0.300	0.319	0.311
preposition	0.201	0.143	0.192	0.226	*0.333	*0.336	*0.344	*0.362
tense	0.128	0.058	0.066	0.058	0.081	0.096	0.089	0.104
lexical choice of noun	0.054	0.054	0.124	0.133	*0.189	*0.216	*0.250	*0.258
lexical choice of verb	0.098	0.098	0.087	0.138	*0.196	*0.232	*0.232	*0.241
pronoun	0.112	0.063	0.131	0.150	0.177	0.195	0.213	0.213
agreement	0.340	0.197	0.224	0.248	0.260	0.284	0.307	0.307
adjective	0.206	0.094	0.165	0.219	*0.413	*0.426	*0.426	*0.446
verb other	0.109	0.204	0.240	0.311	0.291	*0.340	0.308	0.340
adverb	0.333	0.286	0.286	0.302	0.333	0.349	0.349	0.349
conjunction	0.161	0.161	0.161	0.191	0.161	0.191	0.191	0.191
word order	0.048	0.093	0.093	0.091	0.091	0.091	0.091	0.136
noun other	0.200	0.154	0.286	0.286	*0.531	*0.490	*0.490	*0.490
auxiliary verb	0.083	0.160	0.160	0.083	0.083	0.160	0.160	0.160
other lexical choice	0.182	0.000	0.095	0.095	0.400	0.400	0.400	0.400
relative	0.154	0.285	0.154	0.154	0.154	0.154	0.154	0.154
interrogative	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Total	0.148	0.146	0.180	0.200	0.239	0.247	0.254	0.260

冠詞、前置詞、語彙選択はコーパス量が効きます

	learner	correct
article	I like a chocolate very much.	I like _ chocolate very much.
lexical choice of noun	my cycle was injured, but i wasn't.	my bicycle was damaged, but i wasn't.

Table 5: Examples of system output for article and lexical choice of noun error

性能が向上した誤り種類
Article, preposition, lexical choice of noun, lexical choice of verb, adjective, and noun other
→ローカルな情報で直せる、かつ語彙のカバレッジが重要

名詞の数、時制、動詞の一致は フレーズベース統計翻訳では難しい

	learner	correct
noun number 1	I read various type books.	I read various types of books.
*noun number 2	There is a big snoopy dools in my room.	There is a big snoopy doll in my room.
tense 1	If I 'll live in saitama, I must have ...	If I _ live in saitama, I must have ...
*tense 2	The weather is very sunny, so we were ...	The weather was very sunny, so we were ...
agreement 1	Flowers is very beautiful.	Flowers are very beautiful.
*agreement 2	I think, reading comics are not "reading"	I think, reading comics is not "reading"

Table 6: Examples of system results for noun number, tense and agreement errors. Asterisks indicate that the SMT system using full Lang-8 Corpus failed to correct the errors.

- 長距離の依存関係をモデルに入れる必要がある
- 文を超えた文脈を考慮する必要がある

Lang-8コーパスを用いた英作文の誤り訂正のまとめ

- フレーズベース統計的機械翻訳手法を用いて、全ての誤りを対象とした文法誤り訂正システムを構築した。
- フレーズベースSMTは局所的な文脈のみで解け、かつ語彙のカバレッジが問題となる誤りではコーパスサイズが誤り訂正性能向上に貢献。
- 大域的な文脈が必要な誤りに関してはコーパスサイズを増やしても誤り訂正性能は向上しない。

学習者の誤りパターンを考慮した英語学習者の動詞選択支援

澤井悠, 小町守, 松本裕治

ACL 2013 short (to appear)
言語処理学会年次大会 (2013)

動詞選択は英語学習者にとって難しいです

They _____ with other businessmen and do their jobs with the help of computers.

connect　communicate

動詞選択は重要; なぜなら...

動詞は文の意味の中核を担っている

→ 動詞を間違えると意味が変わってしまう

どんな習熟度の学習者にとっても難しい

動詞選択は英語学習者にとって難しいです

CLC-FCE コーパス
(中級レベル) で
2番目に多い誤り

KJ コーパス
(初級レベル) で
6番目に多い誤り

動詞選択は重要; なぜなら...

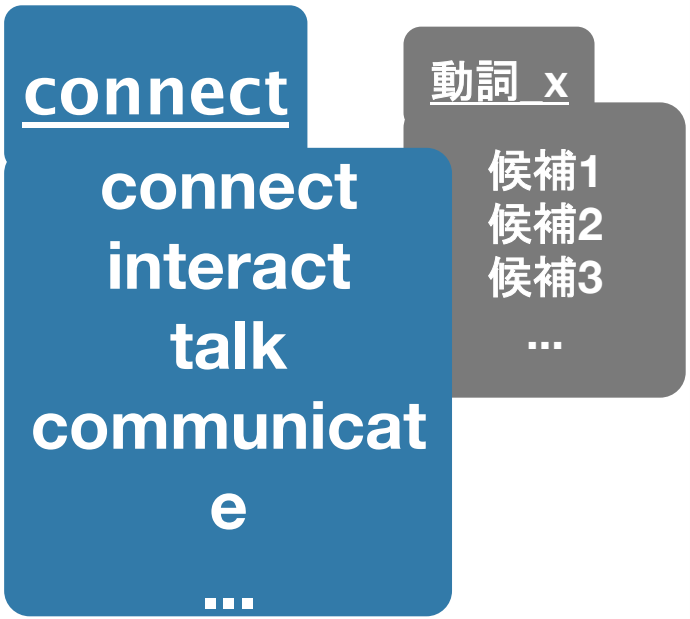
動詞は文の意味の中核を担っている

→ 動詞を間違えると意味が変わってしまう

どんな習熟度の学習者にとっても難しい

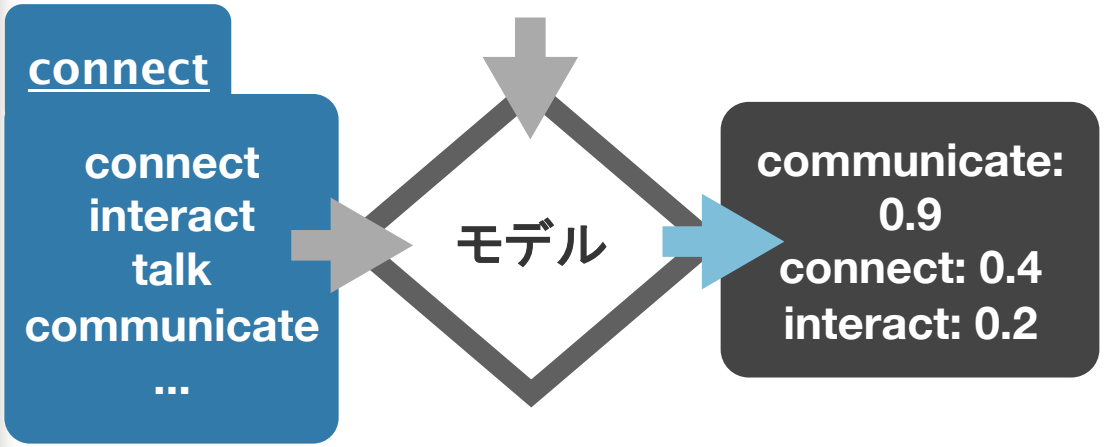
Confusion set と機械学習を用いた動詞選択

“confusion set” の構築



統計モデルによる訂正・推薦

They **connect** with other businessmen...



誤りパターンを考慮した動詞選択手法

タスク:

英語学習者の作文における動詞選択誤り推薦

提案手法と主要な貢献:

大規模学習者コーパスから作成した **confusion set** はカバー率を向上させる

分野適応によって推薦性能が向上する

関連研究

confusion set を用いた動詞誤り推薦・訂正

[Wu+ 2010]: “コロケーション誤りに対する動詞推薦”

confusion set を用いた多クラス分類

[Liu+ 2011]: “中国人英語学習者に対する動詞誤り訂正”

500 動詞に対する **confusion set** を用いた多クラス分類

Jian-Cheng Wu et al. 2010. Automatic Collocation Suggestion in Academic Writing, in Proceedings of ACL, 2010, 115-119.

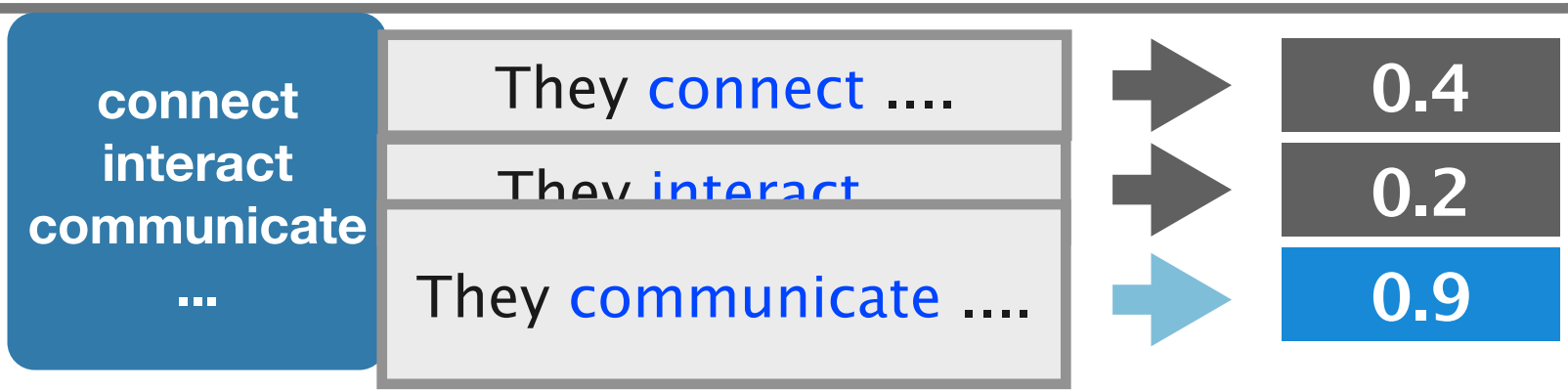
Xiaohua Liu, Bo Han, and Ming Zhou. 2011. Correcting Verb Selection Errors for ESL with the Perceptron, In Proceedings of CiCling, 2011, 411-423.

中国人英語学習者の動詞誤り訂正 [Liu+ 2011]

パーセプトロンを用いた多クラス分類:

confusion set から最尤候補を選択

They **connect** with other businessmen and do their jobs with the help of computers.



Xiaohua Liu, Bo Han, and Ming Zhou. 2011. Correcting Verb Selection Errors for ESL with the Perceptron, in Proceedings of CiCling, 2011, 411–423.

先行研究では誤りパターンを明示的に考慮していません

	Confusion set	モデル
Wu+2010 Liu+2011	学習者の誤りを考慮しない => **低カバー率**	学習者コーパスに適したモデルではない
本研究	大規模学習者コーパスから構築	分野適応によるネイティブコーパスと学習者コーパスで訓練されたモデル

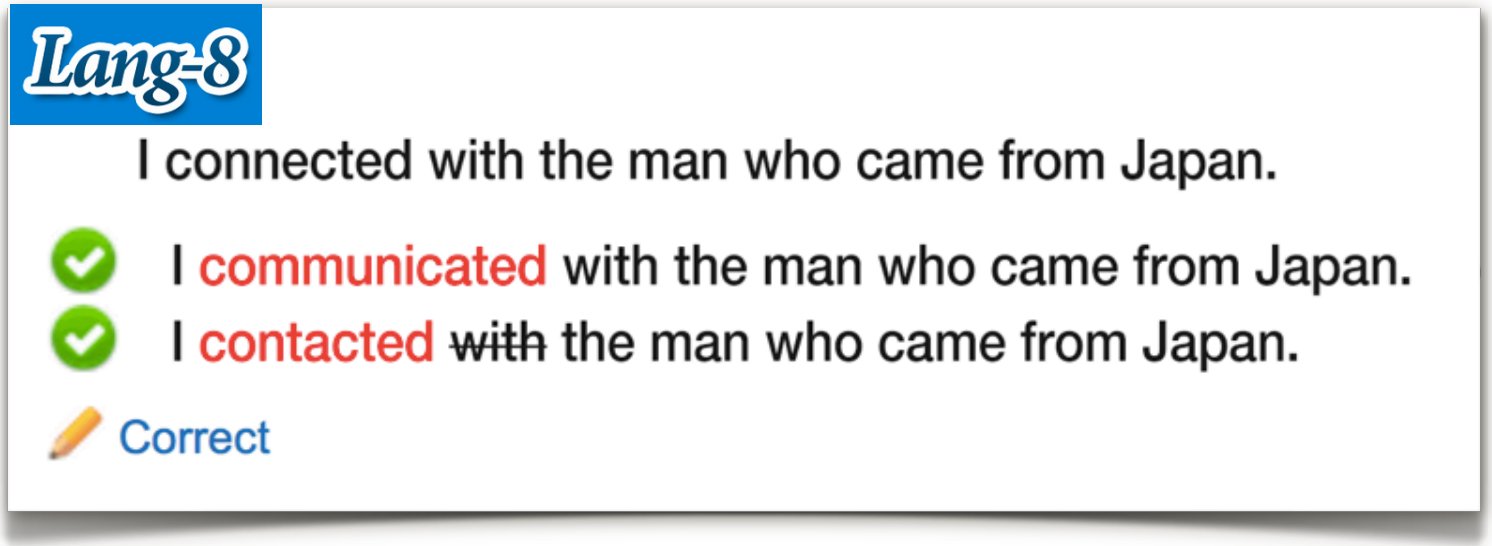
提案手法

confusion set と機械学習

Confusion set のカバー率向上のためにウェブから抽出した知識を用います

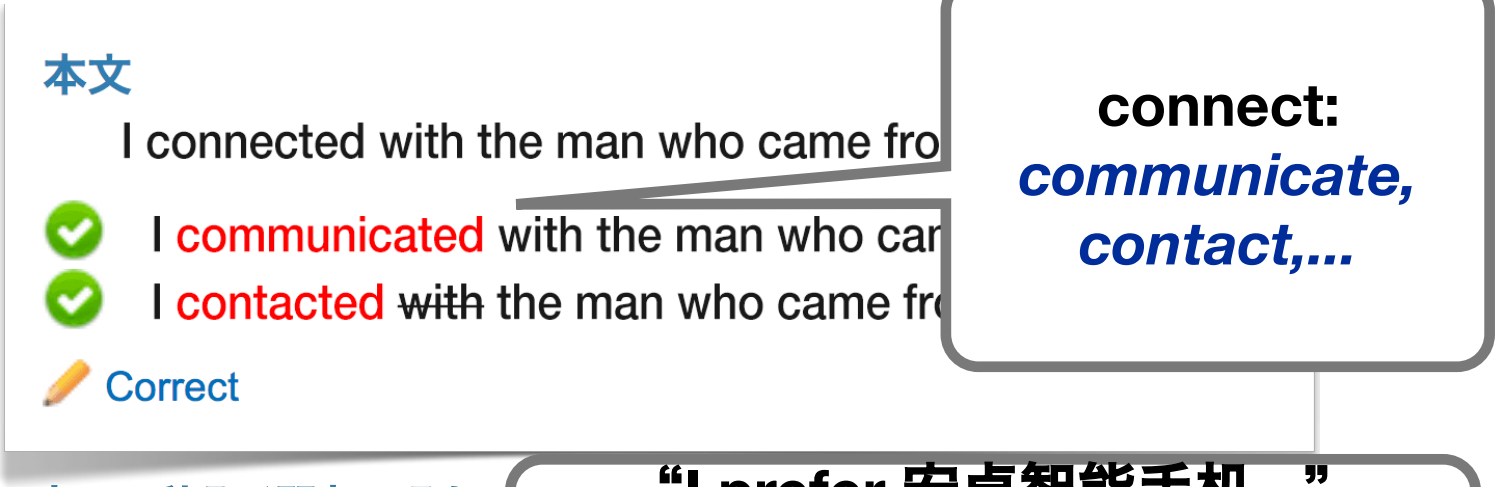
SNS から構築した大規模学習者コーパス

Lang-8から100万文対を取得



Tomoya Mizumoto et al. 2011. Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners, in *Proceedings of IJCNLP2011*

Lang-8 コーパスからの動詞誤りペア抽出



50,000 文の動詞選択誤り

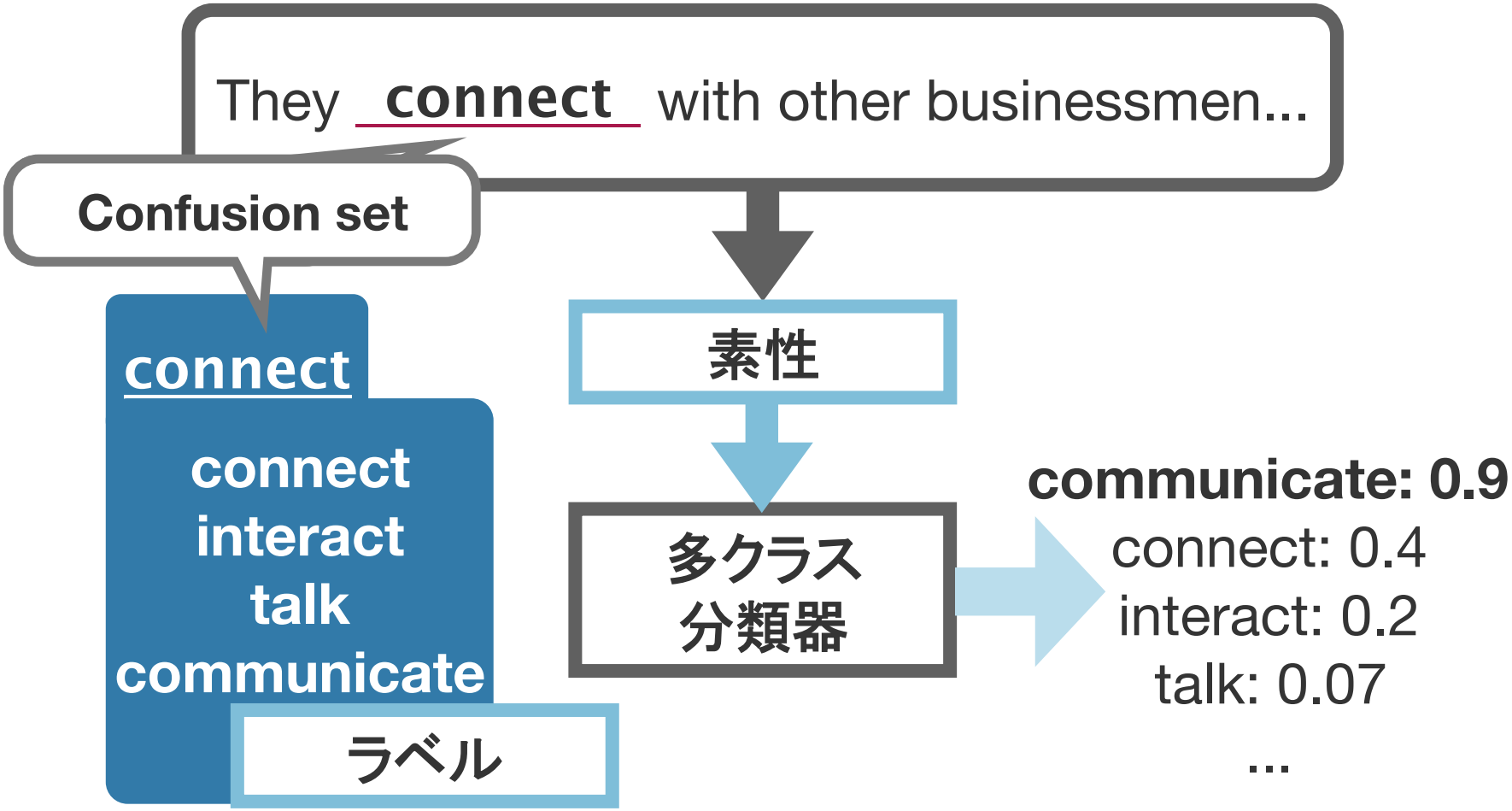
英語でない文対は除外
自動で品詞付与
“動詞の置換” となっている箇所を動的計画法で探索

“I prefer 安卓智能手机...”
“I prefer android smartphones...”

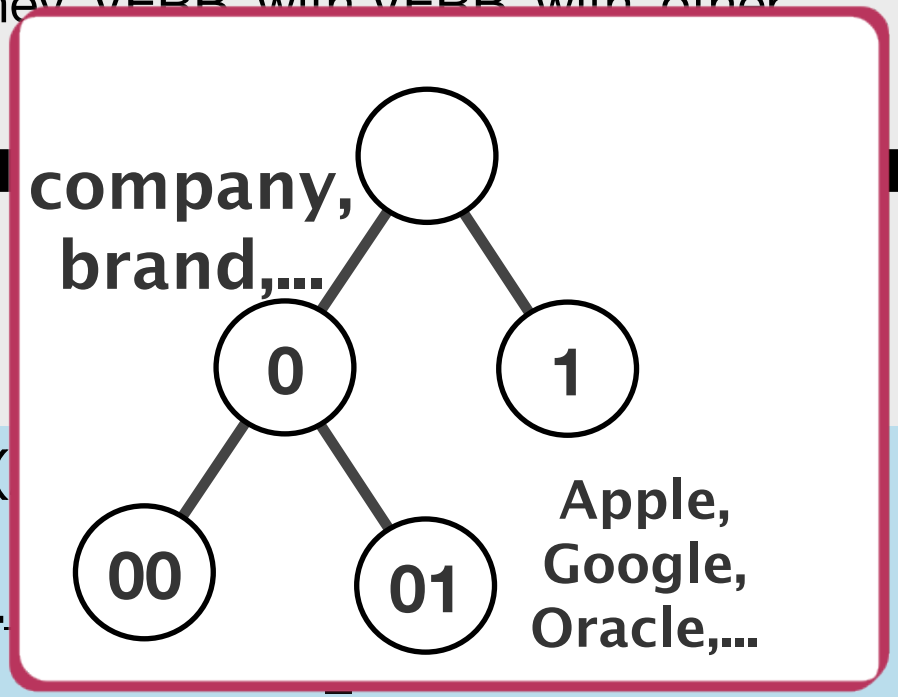
500個の対象動詞に対する Lang-8 の confusion set

10,126 対の動詞誤りパターン

動詞選択を穴埋め問題として定式化



素性	例
3-gram (表層, 品詞) ±2-単語窓	they VERB with VERB with other
名詞句の主辞 (表層, 品詞)	
名詞句の主辞の クラスタ素性 [Brown1992]*	



<s> They _____ with other businessmen and do their jobs with the help of computers. </s>

Peter F Brown et al,. Class-Based N-gram Models of Natural Language, Computational Linguistics 18, no. 4

* https://github.com/percyliang/brown-cluster

分野適応で2つの分野で違う重みを与える

動機: タグ付き学習者コーパスが少ない

元の分野 = ネイティブコーパス, 適応分野 = 学習者コーパス

	素性ベクトル
ネイティブX_S=	[1, 1, 1, 0, 0]
学習者 X_T=	[0, 0, 1, 1, 1]

Hal Daumé III. 2007. Frustratingly Easy Domain Adaptation, in Proceedings of ACL, 2007, 256–263.

EasyAdapt [DaumeIII 2007] による学習者コーパスへの適応

元々の素性を “分野に特化” した素性集合にもマップ

$$\Phi^S = \langle x, x, 0 \rangle$$
$$\Phi^T = \langle x, 0, x \rangle$$

	共通	元の分野	適応分野
ネイティブ X_S' =	[1, 1, 1, 0, 0,	1, 1, 1, 0, 0,	0, 0, 0, 0, 0]
学習者 X_T' =	[0, 0, 1, 1, 1,	0, 0, 0, 0, 0,	0, 0, 1, 1, 1]

Hal Daumé III. 2007. Frustratingly Easy Domain Adaptation, in Proceedings of ACL, 2007, 256-263.

実験

2つの学習者コーパスで推薦実験を行います

2つの英語学習者コーパスで評価

CLC–FCE (試験作文, 中級レベル

KJ コーパス (日記スタイル作文,

設定: チェックする場所は既

対象動詞は CLC–FCE で頻出の **500** 動詞

動詞誤り以外の誤りは訂正済み

CLC–FCE では 1,083 事例
KJ コーパスでは 80 事例

動詞の延べ ~90%
をカバー

ベースラインの confusion set

WordNet

7,226 パターン

Roundtrip

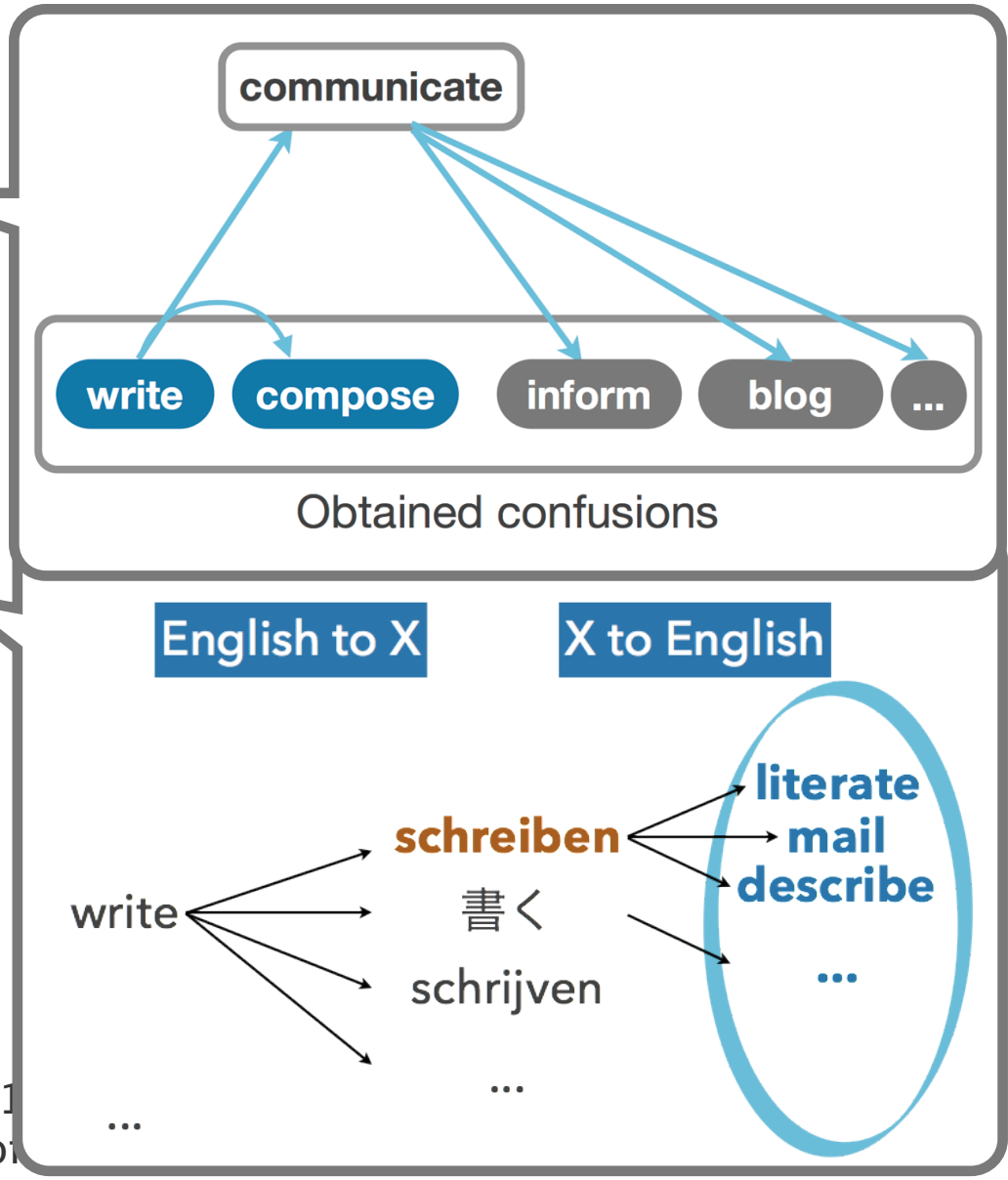
周遊統計的機械翻訳

24,078 パターン

WordNet+Roundtrip:

[Liu+ 2011] と同等,

24,182 パターン



Xiaohua Liu, Bo Han, and Ming Zhou. 2011
ESL with the Perceptron, in Proceedings of

ネイティブコーパスで訓練し、学習者コーパスに適応します

元の分野: ukWaC 英語ネイティブコーパス

さまざまなスタイルで書かれたウェブコ
各動詞に対して 20,000 文 (合計: 1,000

仮定:
CLC-FCE, KJ, Lang-8 は
同じ分野

対象分野: Learner corpora

Lang-8 から抽出した50,000文 (**Lang-8+EasyAdapt**)

学習器: One-vs.-the-rest オンライン線形分類器

L2 正則化 Modified-Huber-loss (scikit-learn 0.14)
全ての実験で同じ素性を使用

平均逆順位 (MRR) によって推薦性能を評価

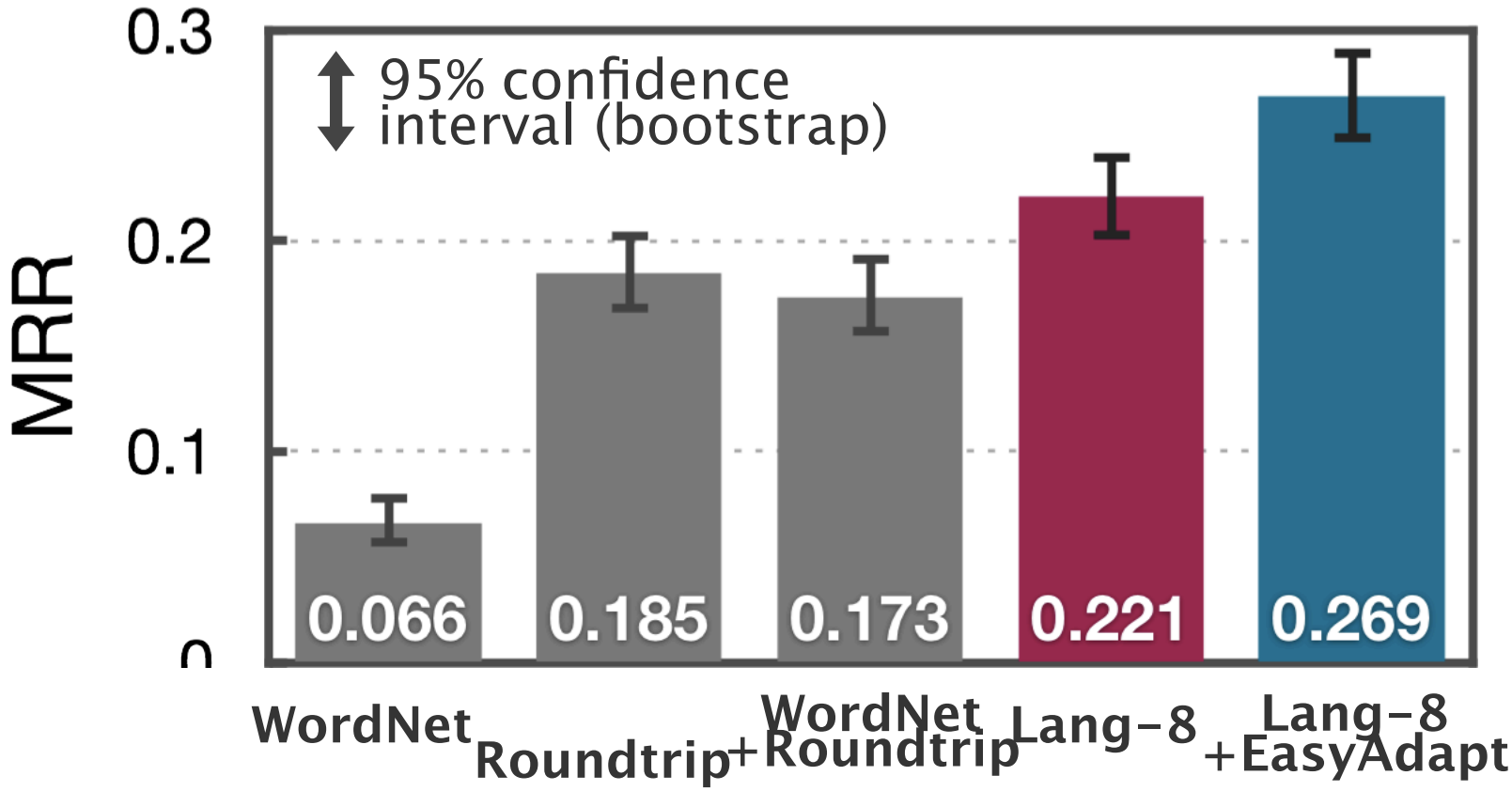
$$MRR = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}\dfrac{1}{rank_gold_i} & (gold_i \in Suggestion_i) \\ 0 & (\text{otherwise})\end{cases}$$

MRR は平均的な推薦の質を反映したものです

値域は [0, 1], 大きいほうがよい

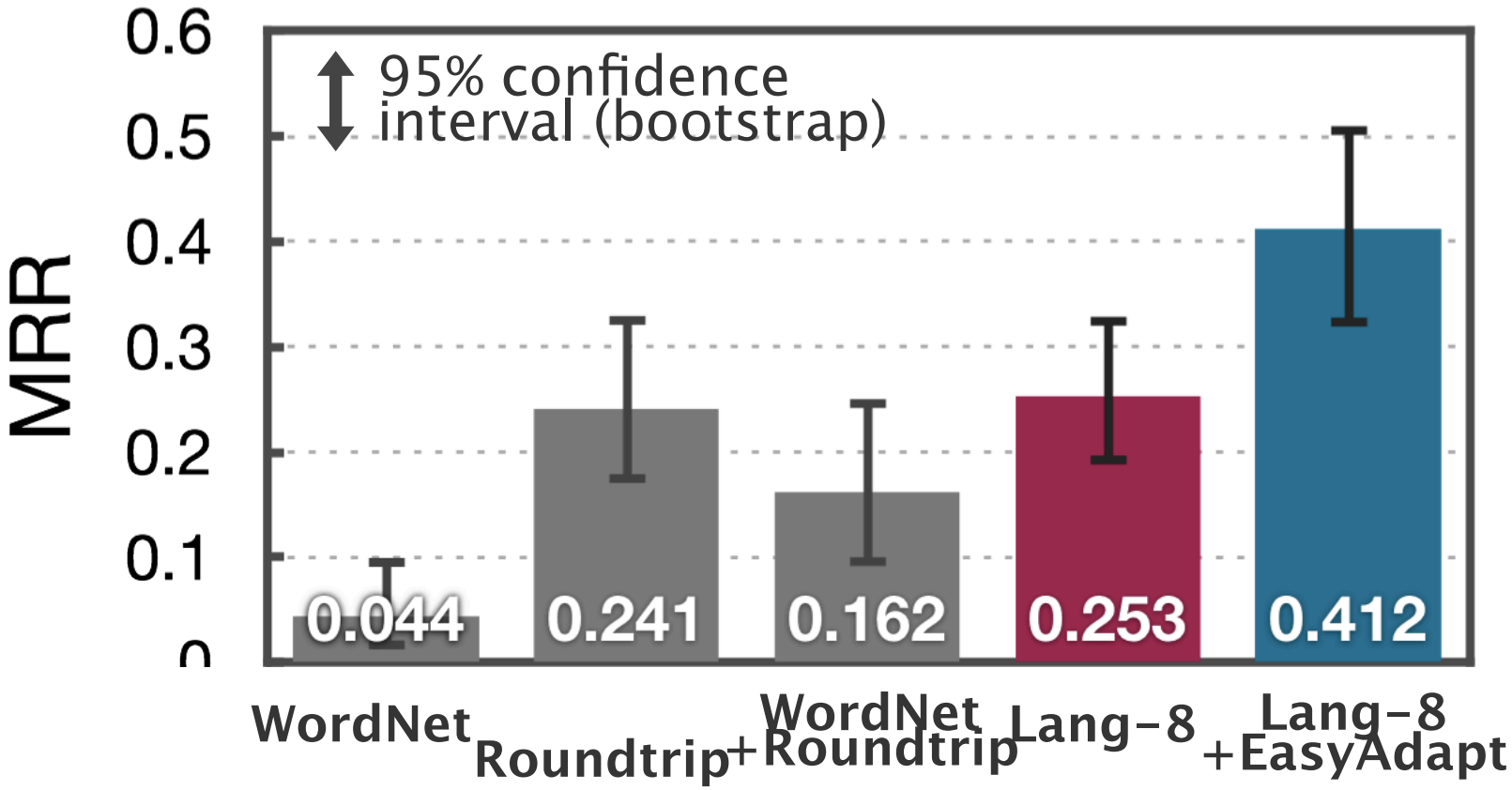
N　　: テスト事例の数
$gold_i$　: i番目のテスト事例に対する正しい候補

CLC-FCE における結果:
提案手法は推薦性能を向上



カバー率	0.14	0.48	0.45	<u>0.58</u>	<u>0.58</u>

KJ コーパスにおける結果:
コーパスによらず提案手法が高カバー率



カバー率	0.05	0.54	0.54	<u>0.69</u>	<u>0.69</u>

議論とエラー分析

学習者コーパスの誤りパターンを考慮した効果 (Lang-8 vs. Lang-8EA)

改善: 280/624 推薦

CLC-FCE

*live/stay, *say/tell, *bring/take, *solve/resolve, ...

同義語・類義語で改善

改善: 33/60 推薦

KJ コーパス

*see/watch, *learn/gain, *tell/teach, *take/talk, ...

よくある誤りで改善
(似た音の動詞を含む)

類義語に対する推薦がうまくいっていません

元の動詞のほうがよかった

47/624 (Lang-8)
30/624 (Lang-8EA)

...it was not as it was ***mentioned/described** in the advertisement I read in the local newspaper.

上位3つの推薦:

問題点:

同じ主語・目的語を持つ同義語・類義語を区別することができない

対応策:

文単位の潜在的な意味素性を用いる

学習者の誤りパターンを考慮した英語学習者の動詞選択支援のまとめ

動詞選択手法を提案

学習者の誤りパターンを明示的に考慮
(1) 学習者コーパスから構築した **confusion set**
(2) 分野適応手法の適用

実際の学習者コーパスを用いた実験結果

学習者コーパスから作成した confusion set によってカバー率が向上し、推薦性能が向上

分野適応によって推薦性能向上

これらの手法は**コーパスに依存しない**

まとめ

- 語学学習 SNS Lang-8 から100万文規模の英語学習者コーパスを抽出した。
- 英語学習者の作文の文法誤り訂正における学習者コーパスの分量の影響を示した。
  - フレーズベース統計翻訳を用いた手法では、局所的な文脈のみで解け、かつ語彙のカバレッジが問題となる誤りではコーパスサイズが誤り訂正性能向上に貢献。
  - 大域的な文脈が必要な誤りに関してはコーパスサイズを増やしても誤り訂正性能は向上しない。
- 学習者の誤りパターンを考慮した英語学習者の動詞選択支援を行った。
  - 学習者コーパスから作成した confusion set によってカバー率が向上し、推薦性能が向上した。