平易なコーパスを用いない テキスト平易化のための 単言語パラレルコーパスの構築

首都大学東京
梶原智之　小町守

テキスト平易化

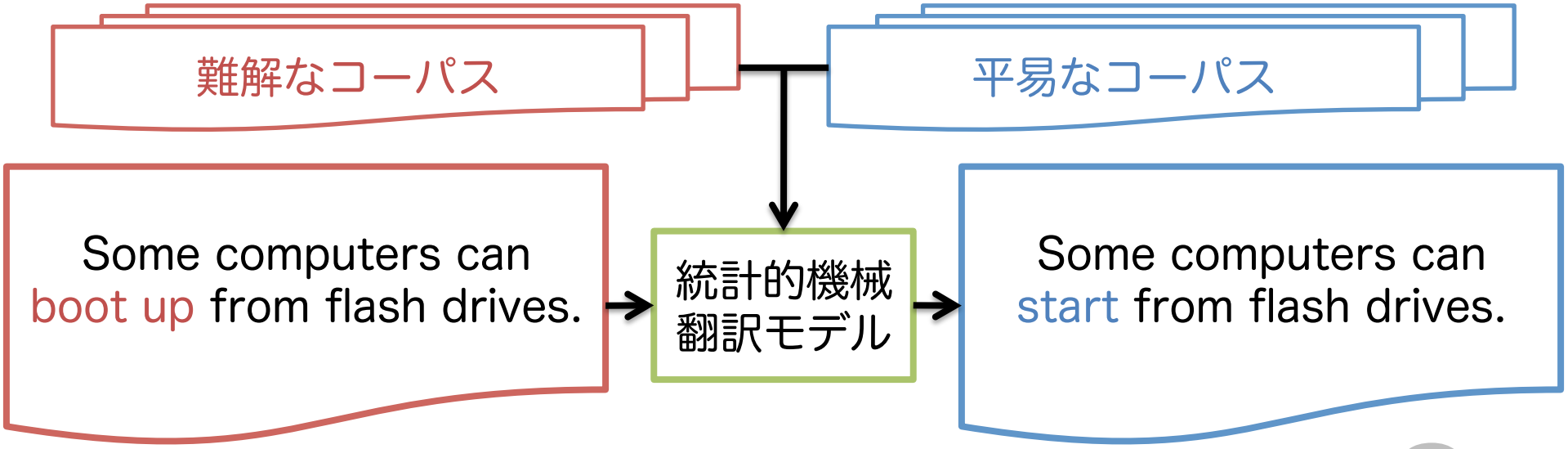
- English Wikipedia: Alfonso Perez
  - Alfonso Perez ~~Munoz, usually referred to as Alfonso,~~ is a former Spanish footballer~~, in the striker position~~.
- Simple English Wikipedia: Alfonso Perez
  - Alfonso Perez is a former Spanish football player.

読みやすくなるように文を書き換えるタスク

- 応用１：自然言語処理のために入力文の複雑さを減らす
- 応用２：言語学習者など人々の文章読解を助ける

統計的機械翻訳の枠組みでのテキスト平易化

- テキスト平易化を同一言語内の翻訳問題と考える
- 難解なテキストと平易なテキストからなる
  パラレルコーパスを用意してトレーニングする



テキスト平易化のための大規模な言語資源は英語でのみ利用可能

英語

- 平易なコーパス：Simple English Wikipedia（100万文）
- 平易なパラレルコーパス：Parallel Wikipedia（50万文対）
  Newsela（5.6万文 × 5段階の難易度）
- 言い換え辞書：PPDB（1億フレーズ対）
- 平易な言い換え辞書：Simple PPDB（450万フレーズ対）

日本語

- 言い換え辞書：PPDB: Japanese（1,500万フレーズ対）

多くの言語で大規模に利用可能な「生コーパス」のみを用いて
テキスト平易化のための言語資源（パラレルコーパス）を構築

平易なコーパスを用いない
テキスト平易化のための単言語パラレルコーパスの構築

Japan is an island country in East Asia.
Japan is a stratovolcanic archipelago of 6,852 islands.
The country is divided into 47 prefectures in eight regions.
Archaeological research indicates that Japan was inhabited as early as the Upper Paleolithic period.
The population of 126 million is the world's tenth largest.

生コーパス

平易なコーパスを用いない
テキスト平易化のための単言語パラレルコーパスの構築

Readability: 80.3	Japan is an island country in East Asia.
Readability: 38.0	Japan is a stratovolcanic archipelago of 6,852 islands.
Readability: 69.8	The country is divided into 47 prefectures in eight regions.
Readability: 15.0	Archaeological research indicates that Japan was inhabited as early as the Upper Paleolithic period.
Readability: 86.7	The population of 126 million is the world's tenth largest.

生コーパス

平易なコーパスを用いない
テキスト平易化のための単言語パラレルコーパスの構築

生コーパス

Readability: 80.3	Japan is an island country in East Asia.
Readability: 38.0	Japan is a stratovolcanic archipelago of 6,852 islands.
Readability: 69.8	The country is divided into 47 prefectures in eight regions.
Readability: 15.0	Archaeological research indicates that Japan was inhabited as early as the Upper Paleolithic period.
Readability: 86.7	The population of 126 million is the world's tenth largest.

難解なコーパス

1. Japan is a stratovolcanic archipelago of 6,852 islands.
2. Archaeological research indicates that Japan was inhabited as early as the Upper Paleolithic period.

平易なコーパス

1. Japan is an island country in East Asia.
2. The country is divided into 47 prefectures in eight regions.
3. The population of 126 million is the world's tenth largest.

① 文のリーダビリティを求めて難解な文と平易な文に分割

平易なコーパスを用いない
テキスト平易化のための単言語パラレルコーパスの構築

生コーパス

Readability: 80.3	Japan is an island country in East Asia.
Readability: 38.0	Japan is a stratovolcanic archipelago of 6,852 islands.
Readability: 69.8	The country is divided into 47 prefectures in eight regions.
Readability: 15.0	Archaeological research indicates that Japan was inhabited as early as the Upper Paleolithic period.
Readability: 86.7	The population of 126 million is the world's tenth largest.

難解なコーパス

1. Japan is a stratovolcanic archipelago of 6,852 islands.
2. Archaeological research indicates that Japan was inhabited as early as the Upper Paleolithic period.

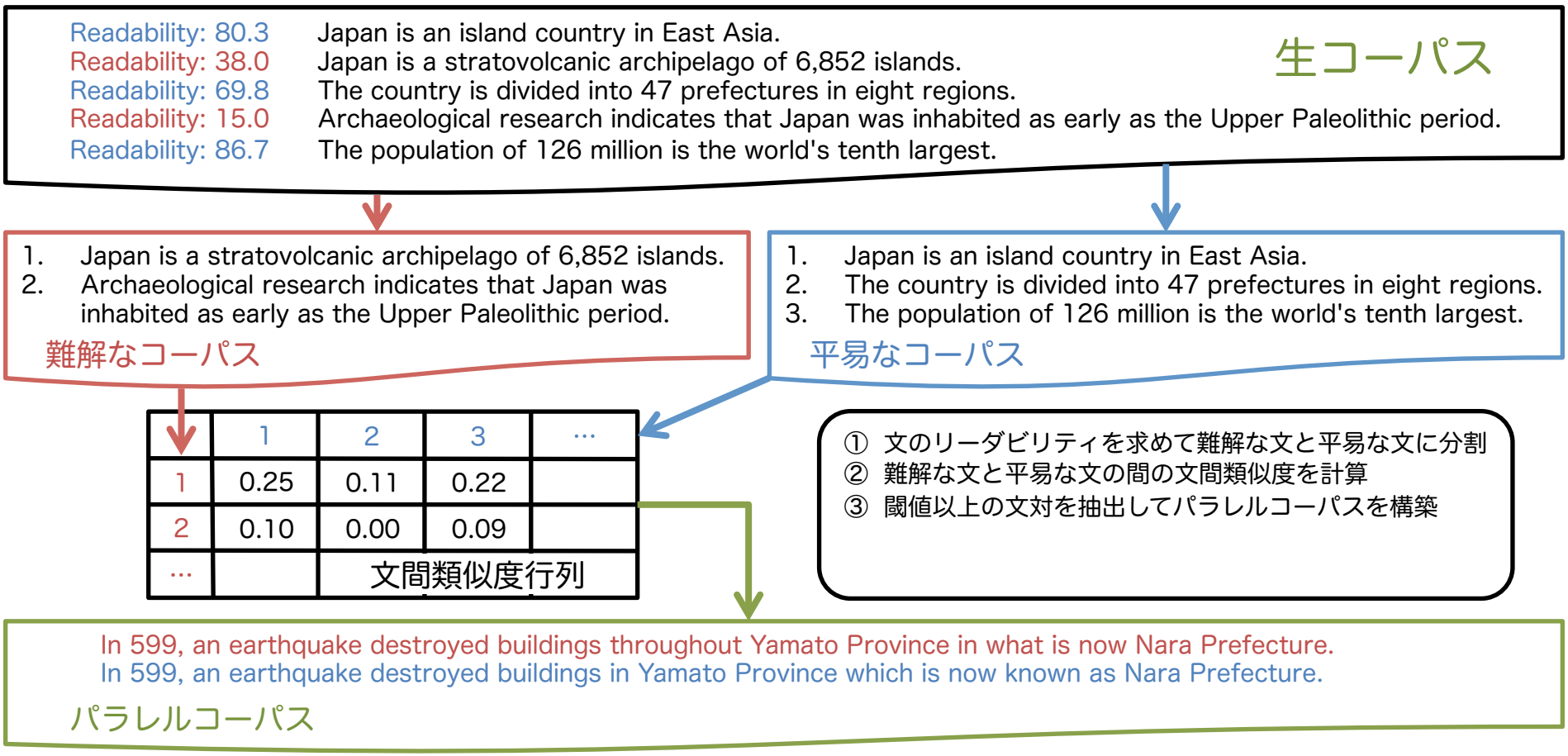
平易なコーパス

1. Japan is an island country in East Asia.
2. The country is divided into 47 prefectures in eight regions.
3. The population of 126 million is the world's tenth largest.

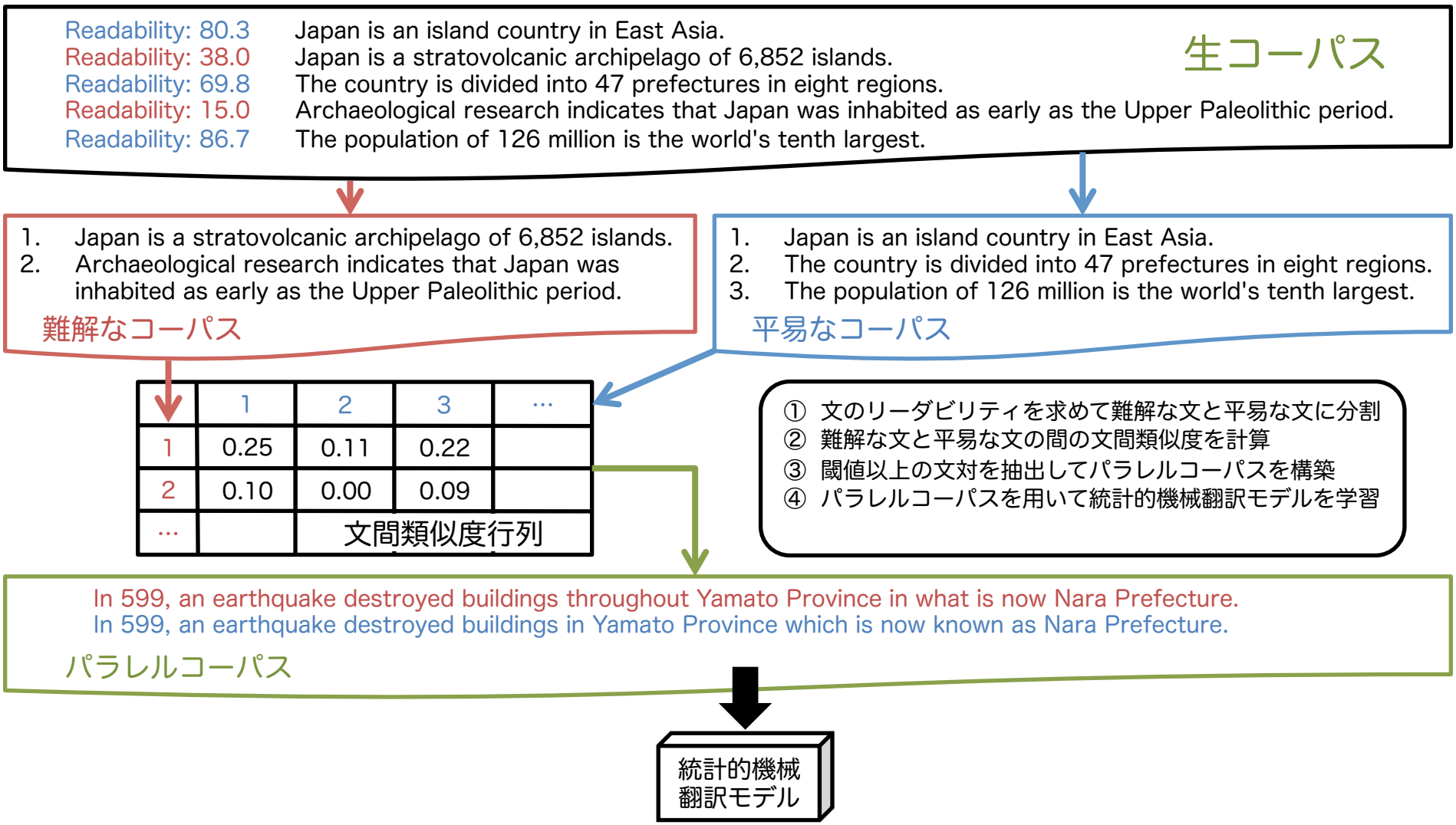
	1	2	3	…
1	0.25	0.11	0.22	
2	0.10	0.00	0.09	
…		文間類似度行列		

① 文のリーダビリティを求めて難解な文と平易な文に分割
② 難解な文と平易な文の間の文間類似度を計算

平易なコーパスを用いない
テキスト平易化のための単言語パラレルコーパスの構築



平易なコーパスを用いない
テキスト平易化のための単言語パラレルコーパスの構築



平易なコーパスを用いない
テキスト平易化のための単言語パラレルコーパスの構築

生コーパス

Readability: 80.3 Japan is an island country in East Asia.
Readability: 38.0 Japan is a stratovolcanic archipelago of 6,852 islands.
Readability: 69.8 The country is divided into 47 prefectures in eight regions.
Readability: 15.0 Archaeological research indicates that Japan was inhabited as early as the Upper Paleolithic period.
Readability: 86.7 The population of 126 million is the world's tenth largest.

難解なコーパス

1. Japan is a stratovolcanic archipelago of 6,852 islands.
2. Archaeological research indicates that Japan was inhabited as early as the Upper Paleolithic period.

平易なコーパス

1. Japan is an island country in East Asia.
2. The country is divided into 47 prefectures in eight regions.
3. The population of 126 million is the world's tenth largest.

	1	2	3	…
1	0.25	0.11	0.22	
2	0.10	0.00	0.09	
…		文間類似度行列		

① 文のリーダビリティを求めて難解な文と平易な文に分割
② 難解な文と平易な文の間の文間類似度を計算
③ 閾値以上の文対を抽出してパラレルコーパスを構築
④ パラレルコーパスを用いて統計的機械翻訳モデルを学習
⑤ モデルを用いて入力文から平易な同義文を生成

パラレルコーパス

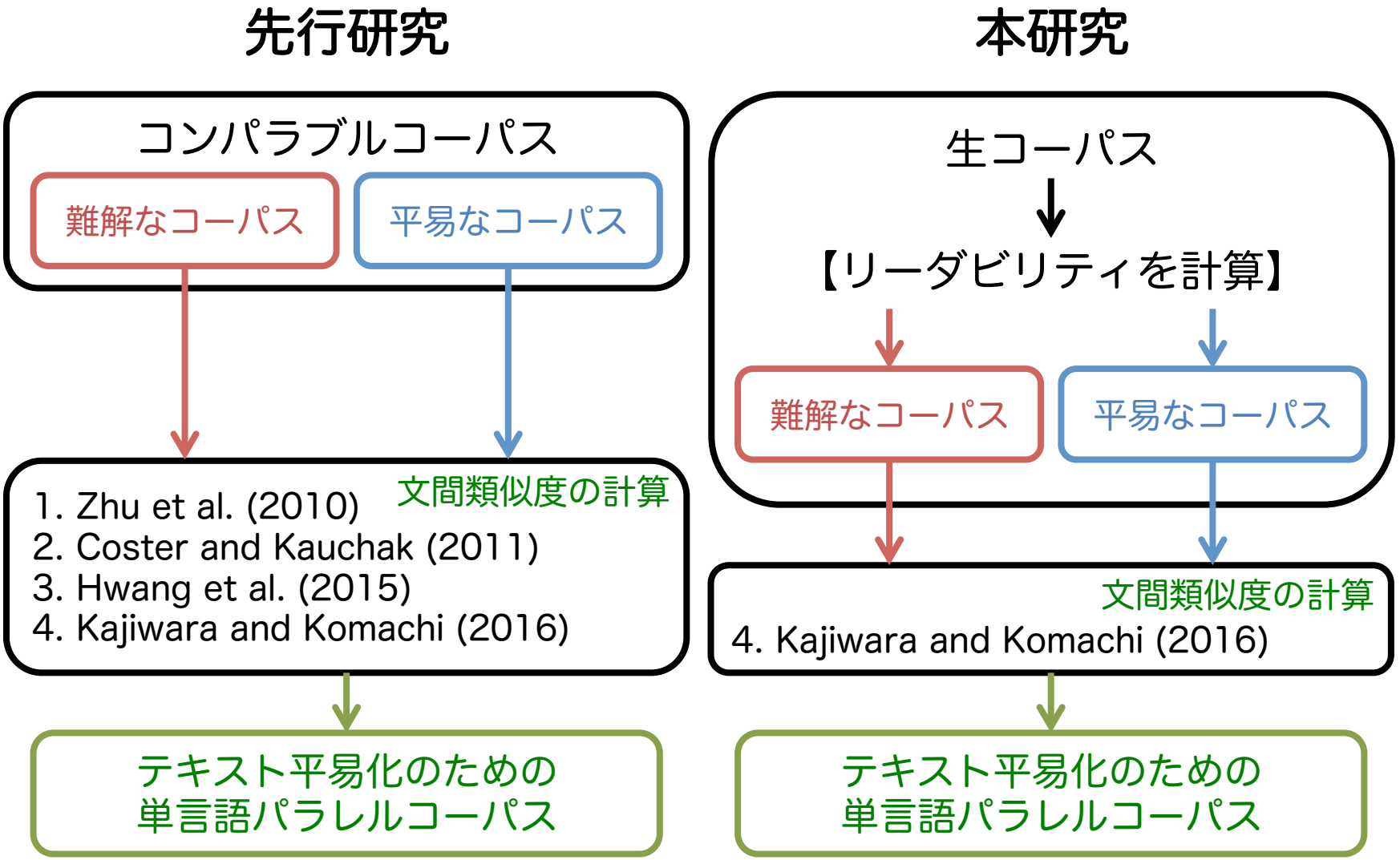
In 599, an earthquake destroyed buildings throughout Yamato Province in what is now Nara Prefecture.
In 599, an earthquake destroyed buildings in Yamato Province which is now known as Nara Prefecture.

Some computers can boot up from flash drives. → 統計的機械翻訳モデル → Some computers can start from flash drives.

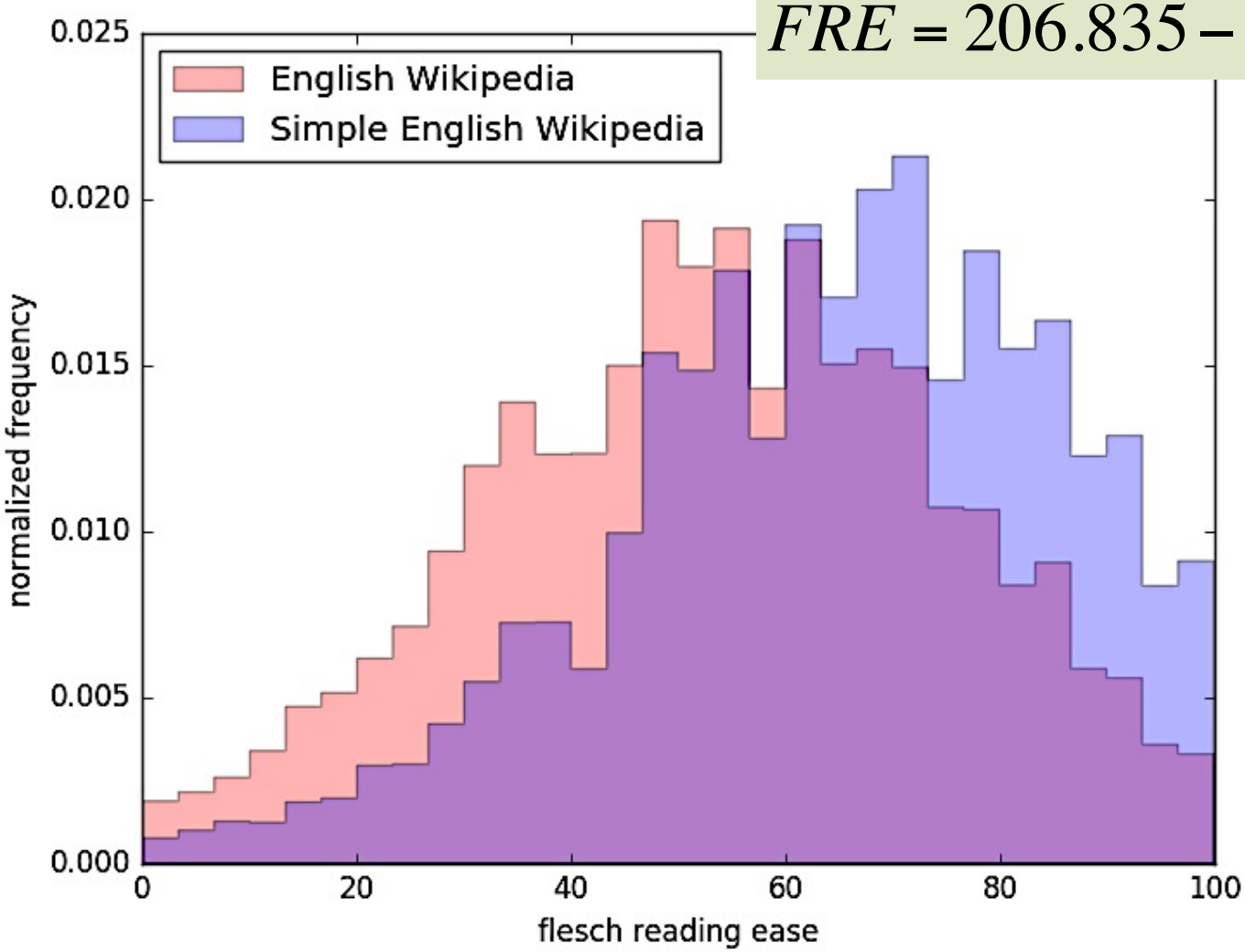
先行研究：英語のテキスト平易化コーパス
（English WikipediaとSimple English Wikipediaから構築）

1. Zhu et al. (2010)　10.8 万文
   - 文をTF-IDFベクトルとして表現
   - ベクトル間のコサイン類似度が閾値を越える文対を抽出
2. Coster and Kauchak (2011)　13.7 万文
   - Zhuらの手法を拡張し、文の出現順序を考慮
3. Hwang et al. (2015)　28.5 万文
   - Wiktionaryの見出し語と定義文中の単語の共起を用いて異なる単語間の類似度を考慮
4. Kajiwara and Komachi (2016)　49.3 万文
   - 単語分散表現を用いて異なる単語間の類似度を考慮

先行研究：文間類似度の計算を工夫
本研究：平易な文の抽出を工夫



リーダビリティに基づく難解な文と平易な文の分類



$$FRE = 206.835 - 1.015\alpha - 84.6\beta$$

α：単語数
β：平均音節数

90 ~ 100	Very Easy
80 ~ 89	Easy
70 ~ 79	Fairly Easy
60 ~ 69	Standard
50 ~ 59	Fairly Difficult
30 ~ 49	Difficult
0 ~ 29	Very Difficult

単語分散表現のアライメントに基づく文間類似度を用いた文アライメント

- 文xyの類似度を、アラインされた単語類似度の平均値で定義
- 単語類似度 $\phi(x_i, y_j)$ にはCBOWベクトルのCOS類似度を使う
- 各単語 x_i に対して、最も類似度が高い単語 y_j をアラインする
- S_{asym} は非対称なスコアなので、両方向の平均値を取る
- ノイズ軽減のため、$\phi <$ (閾値) の単語対はアラインしない

$$S_{asym}(x,y) = \frac{1}{|x|}\sum_{i=1}^{|x|}\max_j \phi(x_i, y_j)$$

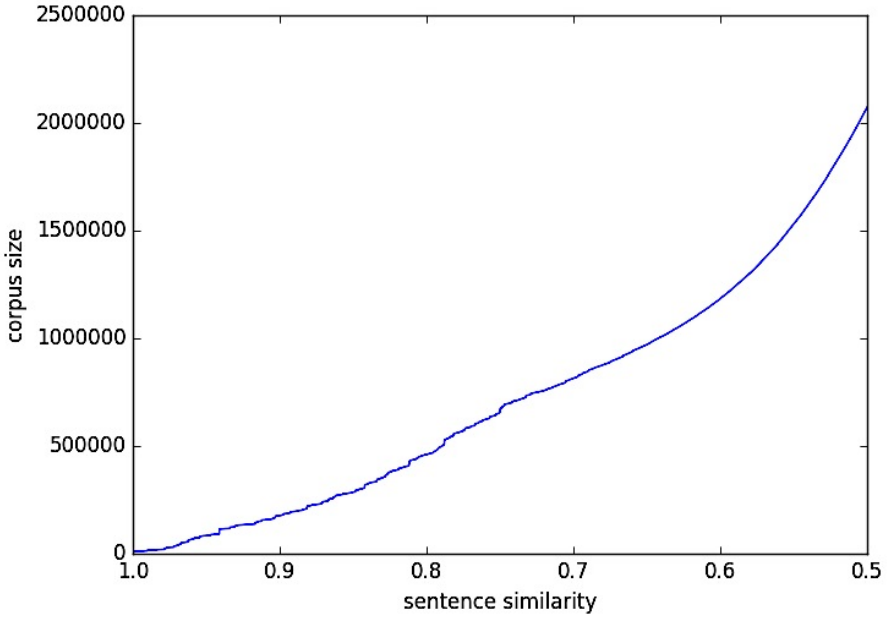
$$S_{sym}(x,y) = \frac{1}{2}(S_{asym}(x,y) + S_{asym}(y,x))$$

Song and Roth (2015) Unsupervised Sparse Vector Densification for Short Text Similarity

200万文対のテキスト平易化コーパス

- English Wikipedia：6,283,703文
  - 難解な（ 0 <= FRE < 60）コーパス：3,689,227文
  - 平易な（60 <= FRE <=100）コーパス：2,358,921文
  - その他（FRE < 0, 100 < FRE）は除外： 235,555文
    ※ 数百単語の長文や箇条書きなど

- 文アライメント
  - 閾値（単語）：$\phi > 0.5$
  - 閾値（文）：$S_{sym} > 0.5$
  - 合計：2,072,572文対



テキスト平易化コーパスの例

類似度	難解な文	平易な文
0.99	Climate in this area has mild differences between highs and lows, and there is adequate precipitation year round.	Climate in this area has mild differences between highs and lows, and there is adequate rainfall year round.
0.88	The new German Empire included 25 states (three of them, Hanseatic cities) ~~and the imperial territory of Alsace-Lorraine.~~	The new German Empire included 25 states, three of them Hanseatic cities.
0.77	In 1996, she received the Primetime Emmy Award for Outstanding Supporting Actress in a Comedy Series, an award she was nominated for on seven occasions.	In 2006 and 2008, she received Emmy nominations for Outstanding Supporting Actress in a Drama Series.
0.66	The album reached number two in the UK Albums Chart and was certified double platinum by the British Phonographic Industry (BPI).	The single reached number one in the UK and has been certified platinum by the BPI, selling 600,000 copies.
0.55	Bombed as a target of the Oil Campaign of World War 2, Erfurt suffered only limited damage and was captured on 12 April 1945, by the US 80th Infantry Division.	During World War 2 the city suffered only some damage and was liberated by the British 8th army on 20 June 1944.

統計的機械翻訳を用いたテキスト平易化

- Moses（PBSMTツール）
- GIZA++（単語アライメント）
- KenLM（言語モデル）
  - 比較手法：Simple English Wikipedia から5-gram
  - 提案手法：リーダビリティ >= 60 の文から5-gram
- マルチリファレンスのテストデータ
  - English Wikipediaから 350文 × 8リファレンス
  - リファレンス：人手で平易に書き換えた文

自動評価尺度

- FRE：リーダビリティを計算する自動評価尺度
  出力のみを用いて評価する
- BLEU：機械翻訳の標準的な自動評価尺度
  意味や文法の観点で人手評価との相関が高い
  出力とリファレンスの2つを比較して評価する
- SARI：テキスト平易化のための自動評価尺度
  難易度も含めてバランス良く人手評価との相関がある
  入力と出力とリファレンスの3つを比較して評価する

$$SARI = \frac{1}{3}F_{add} + \frac{1}{3}F_{keep} + \frac{1}{3}P_{del}$$

Xu et al. (2016) Optimizing Statistical Machine Translation for Text Simplification

平易なコーパスでトレーニングした先行研究と同等の性能を達成

	コーパスサイズ	FRE	BLEU	SARI
Baseline（入力文を書き換えずに出力）	0	54.5	99.4	25.9
Zhu+ の平易なパラレルコーパス	108,016	59.7	84.7	34.7
Coster+ の平易なパラレルコーパス	137,362	59.8	86.4	34.1
Hwang+ の平易なパラレルコーパス	284,738	61.0	81.3	34.5
Kajiwara+ の平易なパラレルコーパス	492,993	61.7	78.4	34.9
Xu et al. (2016)：with PPDB	(tuning) 2,000	67.9	72.4	37.9
本研究：without 平易なコーパス	100,000	54.9	94.9	29.1
本研究：without 平易なコーパス	500,000	55.3	92.7	31.1
本研究：without 平易なコーパス	1,000,000	56.9	88.0	33.7
本研究：without 平易なコーパス	1,500,000	58.2	83.2	**34.4**
本研究：without 平易なコーパス	2,000,000	59.2	79.1	34.1
本研究：without 平易なコーパス	2,072,572	58.9	78.0	34.0

平易なコーパスでトレーニングした先行研究と同等の性能を達成

Input	Offenbach's numerous operettas, such as *Orpheus in the Underworld*, and *La belle Hélène*, were extremely popular in both France and the English-speaking world during the 1850s and 1860s.
Ref 1	Offenbach's ~~numerous~~ operettas, such as *Orpheus in the Underworld*, and *La belle Hélène*, were ~~extremely~~ very popular in ~~both~~ France and the English-speaking world during the 1850's and 1860's.
Ref 2	Offenbach's ~~numerous~~ many operettas, such as *Orpheus in the Underworld*, and *La belle Hélène*, were ~~extremely~~ very popular in ~~both~~ France and in the English-speaking world during the 1850s and 1860s.
Kajiwara+ 2016	Offenbach's ~~numerous~~ many operettas, such as *Orpheus in the Underworld*, and *La belle Hélène*, were ~~extremely~~ very popular in both France and the English-speaking world during the 1850s and 1860s.
Xu+ 2016	Offenbach's ~~numerous~~ many operettas, such as *Orpheus in the Underworld*, and ~~*La*~~ *The belle Hélène*, were ~~extremely~~ very popular in both France and the English~~-speaking~~ world ~~during~~ in the 1850s and 1860s.
本研究	Offenbach's ~~numerous~~ many operettas, such as *Orpheus in the Underworld*, and *La belle Hélène*, were extremely popular in ~~both~~ France and the English-speaking world during the 1850s and 1860s.

なぜ上手く動くのか？

生コーパスを分割して得た難解なコーパスと平易なコーパスは、コンパラブルコーパスではないが、実験結果はこれが問題ではないことを示している

1. PBSMTではフレーズ単位の変換対を学習するから
   - 単語やフレーズの部分的な対応は、同義や含意の関係にある文対からだけではなく、類義の関係にある文対からも得ることができる
2. 言語モデルでのリランキングを行うから
   - 獲得したフレーズ対に雑音が多くても、その中に適切なフレーズ対を含むことができれば流暢な平易文を生成することができる

まとめと今後

- 生コーパスのみを用いてテキスト平易化のための単言語パラレルコーパスを構築した
- PBSMTを用いたテキスト平易化の実験によって、平易な大規模コーパスを用いてトレーニングする場合と同等の成果を得られることがわかった
- 生コーパスは英語以外の言語でも大規模に利用できる
- 今後は任意の言語でテキスト平易化が実現できるだろう（ソースコードをGitHubで公開予定）