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

（第227回自然言語処理研究会の発表 +α）

首都大学東京

梶原智之　小町守

自己紹介：梶原智之（かじわらともゆき）

- 新居浜工業高等専門学校（2006年4月～2011年3月）
  - 電気情報工学科
  - 音楽情報処理（遺伝的アルゴリズムを用いた自動作曲）
- 長岡技術科学大学（2011年4月～2015年3月）
  - 修士課程：工学研究科 電気電子情報工学専攻
  - 自然言語処理（文章読解支援のための語彙平易化）
- 首都大学東京（2015年4月～）　※ 博士後期課程2年目
  - 博士後期課程：システムデザイン研究科 情報通信システム学域
  - 自然言語処理（文章読解支援のためのテキスト平易化）
  - NLP若手の会プログラム委員, NLP東京Dの会（主催）
  - https://sites.google.com/site/moguranosenshi/

修士：日本語の語彙平易化システム

難解な日本語（入力文）
未来は若者が担う

↓

難解語の検出
担う
形態素解析 + 平易語リスト

↓

言い換え
担う：伝承する, 支える, 受け継ぐ
分布類似度,国語辞典,対訳コーパス

→

文脈に合わない語の削除
担う, 支える, 受け継ぐ
述語項構造解析 + 格フレーム辞書

↑

平易な順にランキング
1: 支える，2: 受け継ぐ，3: 担う
N-gram頻度,単語親密度,使用者数

↑

平易な日本語（出力文）
未来は若者が支える

- NLP若手の会第7回シンポジウム奨励賞
- 情報処理学会第77回全国大会学生奨励賞
- 長岡技術科学大学電気系系長賞

- 横浜市：日本語非母語話者のために公的文書を書き換える
- 東京都：東京オリンピック？

修士：日本語の語彙平易化システム

【百貨店】から離れがちな【顧客】を、どう引き戻すか。
【デパート】から離れがちな【お客さん】を、どう引き戻すか。

【よもや】と思う変化が【いとも】簡単に起こる。
【まさか】と思う変化が【とても】簡単に起こる。

自覚の【欠如】が【嘆かわしい】。
自覚の【不足】が【悲しい】。

その笑顔には、子供を【慈しむ】父親の【眼差し】があった。
その笑顔には、子供を【愛する】父親の【視線】があった。

【ただただ】【感嘆する】ばかりである。
【とにかく】【感動する】ばかりである。

自己紹介：梶原智之（かじわらともゆき）

- 新居浜工業高等専門学校（2006年4月～2011年3月）
  - 電気情報工学科
  - 音楽情報処理（遺伝的アルゴリズムを用いた自動作曲）
- 長岡技術科学大学（2011年4月～2015年3月）
  - 修士課程：工学研究科 電気電子情報工学専攻
  - 自然言語処理（文章読解支援のための語彙平易化）
- 首都大学東京（2015年4月～）　※ 博士後期課程2年目
  - 博士後期課程：システムデザイン研究科 情報通信システム学域
  - 自然言語処理（文章読解支援のためのテキスト平易化）
  - NLP若手の会プログラム委員, NLP東京Dの会（主催）
  - https://sites.google.com/site/moguranosenshi/

テキスト平易化

- 難解なテキストの意味を保持したまま平易に書き換える

English Wikipedia: Alfonso Perez

Alfonso Perez ~~Munoz, usually referred to as Alfonso,~~ is a former Spanish footballer, ~~in the striker position.~~

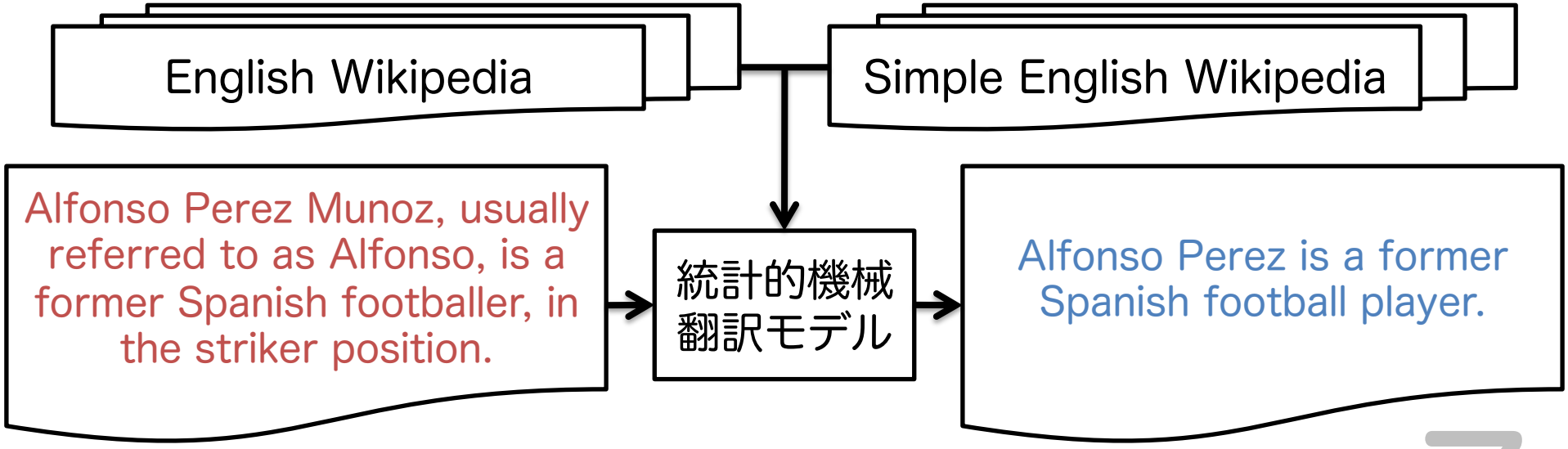
Simple English Wikipedia: Alfonso Perez

Alfonso Perez is a former Spanish football player.

- 文圧縮 + 言い換え
- テキスト平易化は言語学習者や子どもを
  はじめとする多くの読者の文章読解を支援する

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いて統計的機械翻訳モデルを学習



- 英語、ポルトガル語、スペイン語、日本語など

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

1. Lennon was born in war-time England, on 9 October 1940 at Liverpool Maternity Hospital, to Julia and Alfred Lennon, a merchant seaman of Irish descent, who was away at the time of his son‘s birth.
2. His parents named him John Winston Lennon after his paternal grandfather, John “Jack” Lennon, and then-Prime Minister Winston Churchill. …

難解なコーパス

1. Lennon started the Beatles in his hometown of Liverpool, with Paul McCartney and George Harrison.
2. After Ringo Starr joined the band, they started to be very successful.
3. People were excited by their music, and their live performances always pleased audiences. …

平易なコーパス

	1	2	3	…
1	0.27	0.10	0.05	
2	0.19	0.01	0.07	
…		文間類似度行列		

"Lucy in the Sky with Diamonds" is a song written primarily by John Lennon and credited to Lennon–McCartney, for the Beatles' 1967 album Sgt. Pepper's Lonely Hearts Club Band.
"Lucy in the Sky with Diamonds" is a song written by John Lennon and Paul McCartney for The Beatles' 1967 album Sgt. Pepper's Lonely Hearts Club Band. (0.91)

After his marriage to Yoko Ono in 1969, he changed his name to John Ono Lennon.
Lennon loved his wife so much that he added her surname Ono to his own name, since she became Yoko Ono Lennon when she married him. (0.53)

パラレルコーパス

① 単語分散表現のアライメントに基づく文間類似度の計算
② 閾値以上の文対を抽出してパラレルコーパスを構築
③ 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
④ モデルを用いて入力文から平易な同義文を生成

John Lennon was an English singer and songwriter who rose to worldwide fame as a co-founder of the Beatles, the most commercially successful band in the history of popular music.

統計的機械翻訳モデル

John Lennon was an English singer, songwriter and artist who rose to worldwide fame as the founder of the rock band the Beatles.

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

1. Lennon was born in war-time England, on 9 October 1940 at Liverpool Maternity Hospital, to Julia and Alfred Lennon, a merchant seaman of Irish descent, who was away at the time of his son's birth.
2. His parents named him John Winston Lennon after his paternal grandfather, John "Jack" Lennon, and then-Prime Minister Winston Churchill. …

難解なコーパス

1. Lennon started the Beatles in his hometown of Liverpool, with Paul McCartney and George Harrison.
2. After Ringo Starr joined the band, they started to be very successful.
3. People were excited by their music, and their live performances always pleased audiences. …

平易なコーパス

	1	2	3	…
1	0.27	0.10	0.05	
2	0.19	0.01	0.07	
…		文間類似度行列		

難解な文と平易な文に対して、分散表現を用いた多対一の単語アライメントを考え、それらの単語間類似度の平均値を用いて文間類似度を計算する

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

内的評価：文間類似度を用いてパラレルとノンパラレルの２値分類を行いＦ値を比較

	1	2	3	…
1	0.27	0.10	0.05	
2	0.19	0.01	0.07	
…		文間類似度行列		

"Lucy in the Sky with Diamonds" is a song written primarily by John Lennon and credited to Lennon–McCartney, for the Beatles' 1967 album Sgt. Pepper's Lonely Hearts Club Band.
"Lucy in the Sky with Diamonds" is a song written by John Lennon and Paul McCartney for The Beatles' 1967 album Sgt. Pepper's Lonely Hearts Club Band. (0.91)

After his marriage to Yoko Ono in 1969, he changed his name to John Ono Lennon.
Lennon loved his wife so much that he added her surname Ono to his own name, since she became Yoko Ono Lennon when she married him. (0.53)

パラレルコーパス

先行研究よりもF値を3.1改善（0.607→0.638）

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

外的評価：パラレルコーパスから統計的機械翻訳モデルを学習し、BLEUを比較

先行研究よりもBLEUを3.2改善（44.3→47.5）

"Lucy in the Sky with Diamonds" is a song written primarily by John Lennon and credited to Lennon–McCartney, for the Beatles' 1967 album Sgt. Pepper's Lonely Hearts Club Band.
"Lucy in the Sky with Diamonds" is a song written by John Lennon and Paul McCartney for The Beatles' 1967 album Sgt. Pepper's Lonely Hearts Club Band. (0.91)

After his marriage to Yoko Ono in 1969, he changed his name to John Ono Lennon.
Lennon loved his wife so much that he added her surname Ono to his own name, since she became Yoko Ono Lennon when she married him. (0.53)

パラレルコーパス

John Lennon was an English singer and songwriter who rose to worldwide fame as a co-founder of the Beatles, the most commercially successful band in the history of popular music.

→ 統計的機械翻訳モデル →

John Lennon was an English singer, songwriter and artist who rose to worldwide fame as the founder of the rock band the Beatles.

英語のテキスト平易化コーパス

- Zhu et al. (2010)
  - 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
- Coster and Kauchak (2011)
  - Zhu et al. (2010) を拡張し、文の出現順序を考慮

異なる単語間（難解／平易）の類似度を考慮したい

- Hwang et al. (2015)
  - 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
- 本研究
  - 単語分散表現を用いて異なる単語間の類似度を考慮

単語分散表現のアライメントに基づく文間類似度の計算

単語分散表現を用いることで、ラベル付きデータや辞書などの外部知識に頼らずに、異なる単語間の類似度を考慮した文間類似度を計算する４手法を応用

Song and Roth (2015) の文間類似度計算手法

1. Average Alignment （多対多の単語アライメント）
2. Maximum Alignment （多対一の単語アライメント）
3. Hungarian Alignment （一対一の単語アライメント）

Kusner et al. (2015) の文間類似度計算手法

4. Word Mover’s Distance （多対多の単語アライメント）

1. Average Alignment

- 文xと文yの間の全ての単語ペアの単語間類似度を計算
- |x||y|個の単語間類似度を平均して文間類似度を求める

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

- x_i：文xに含まれるi番目の単語
- y_j：文yに含まれるj番目の単語
- Φ（x_i, y_j）：単語x_iと単語y_jの間の単語間類似度
  本研究ではコサイン類似度を用いる

2. Maximum Alignment

- Average Alignmentは直感的であるが、多くの単語間類似度はゼロに近い値を取るノイズとなる
- そこで、各単語x_iに対して最も類似度が高い単語y_jのみを用いて文間類似度を計算する

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

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

- $S_{asym}(x,y)$と$S_{asym}(y,x)$を平均して対称な類似度を得る

3. Hungarian Alignment

- 次に一対一の単語アライメントに基づく文間類似度を計算
  - Average Alignment：多対多の単語アライメント
  - Maximum Alignment：多対一の単語アライメント
- 文xと文yを、単語をノード、単語間類似度をエッジとする重み付き完全2部グラフと考える
- このグラフの最大マッチングを求めると、単語間類似度の総和を最大化する一対一の単語アライメントが得られる
- 2部グラフの最大マッチング問題はHungarian法で解ける

$$S_{hun}(x,y) = \frac{1}{\min(|x|,|y|)} \sum_{i=1}^{|x|} \phi(x_i, h(x_i))$$

4. Word Mover's Distance

- Earth Mover's Distanceの特殊な場合に相当する文xから文yへと単語を輸送する輸送問題を解くWMDも多対多の単語アライメントに基づく文間類似度の計算に応用できる

$$S_{wmd}(x,y) = 1 - WMD(x,y)$$

$$WMD(x,y) = \min \sum_{u=1}^{n} \sum_{v=1}^{n} \mathrm{A}_{uv} \varphi(x_u, y_v)$$

$$\sum_{v=1}^{n} \mathrm{A}_{uv} = \frac{1}{|x|} freq(x_u), \sum_{u=1}^{n} \mathrm{A}_{uv} = \frac{1}{|y|} freq(y_v)$$

- ψ（x_u, y_v）：単語x_uと単語y_vの間の単語間非類似度（距離）
- freq(x_u)：文x中での単語x_uの出現頻度
- n：語彙数、　A_{uv}：単語の輸送量を表す行列

実験：テキスト平易化コーパス構築

- 内的評価
  - English WikipediaとSimple English Wikipediaから抽出した各文対について、様々な文間類似度を用いてパラレルとノンパラレルの２値分類のF値を比較する
- 外的評価
  - 本研究で構築するテキスト平易化コーパスと既存のテキスト平易化コーパスのそれぞれで統計的機械翻訳モデルを学習し、English Wikipediaの文をSimple English Wikipediaの文へ翻訳するBLEUを比較する

実験設定：内的評価

- 文間類似度を用いたパラレルとノンパラレルの２値分類
- 評価用データセット：Hwang et al. (2015)
  - English WikipediaとSimple English Wikipediaから抽出した67,853文対に4種類のラベルを人手で付与
    - Good (G) ：両方向含意 277文対
    - Good Partial (GP)：片方向含意 281文対
    - Partial (P) ：関係ある 117文対
    - Bad (B) ：関係ない 67,178文対
  - F1の最大値（MaxF1）とPR曲線（AUC）で評価
  - ２つの設定で評価：G vs. Others
    G+GP vs. Others

ラベルについて

- Good (G) 両方向含意 277文対
  - Apple sauce or applesauce is a puree made of apples.
  - Apple sauce (or applesauce) is a sauce that is made from stewed or mashed apples.
- Good Partial (GP) 片方向含意 281文対
  - Commercial versions of applesauce are really available in supermarkets.
  - It is easy to make at home, and it is also sold already made in supermarkets as a common food.
- Partial (P) 関係ある 117文対
  - Applesauce is a sauce that is made from stewed and mashed apples.
  - Applesauce is made by cooking down apples with water or apple cider to the desired level.
- Bad (B) 関係ない 67,178文対
  - Commercial versions of applesauce are really available in supermarkets.
  - Peeled or unpeeled apples can be used and different spices or additives like cinnamon can be used.

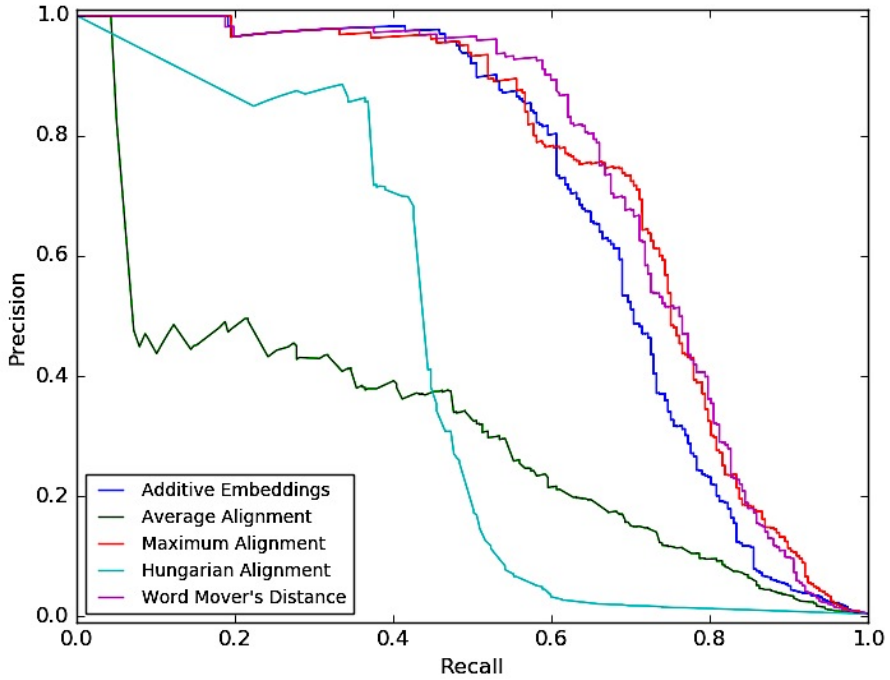
Maximum Alignment が優秀

	G vs. O		G+GP vs. O	
	MaxF1	AUC	MaxF1	AUC
Zhu et al. (2010)	0.550	0.509	0.431	0.391
Coster and Kauchak (2011)	0.564	0.495	0.415	0.387
Hwang et al. (2015)	0.712	0.694	0.607	0.529
Skip-Thought Vectors	0.631	0.544	0.442	0.337
Additive Embeddings	0.691	0.695	0.518	0.487
1. Average Alignment	0.419	0.312	0.391	0.297
2. Maximum Alignment	0.717	0.730	0.638	0.618
3. Hungarian Alignment	0.524	0.414	0.354	0.275
4. Word Mover's Distance	0.724	0.738	0.531	0.499

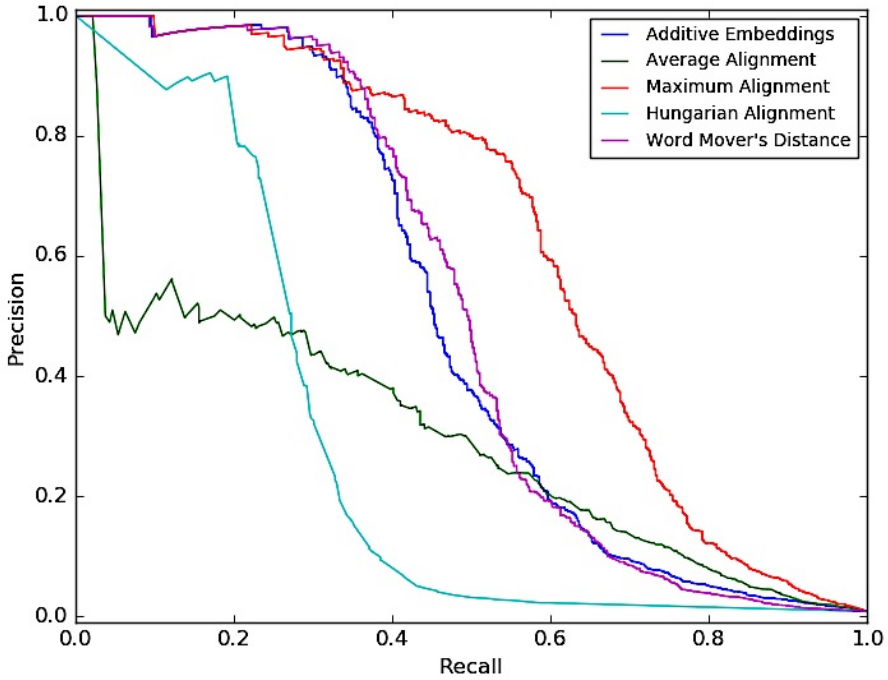
※ Skip-Thought Vectors：RNNで前後の文を予測する文ベクトルのcos類似度
※ Additive Embeddings：単語分散表現を足した文ベクトルのcos類似度

Maximum Alignment が優秀

G vs. O



G+GP vs. O



- テキスト平易化では、重要ではない難解な表現の省略によって読みやすい短文を生成することも多い。
- そのため、難解な文と平易な文が同義であるGの文対だけでなく、難解な文が平易な文を含意するGPの文対もコーパスに含めることが重要。

テキスト平易化コーパスの構築

- English WikipediaとSimple English Wikipediaから
  タイトルが一致する 126,725 文書対を収集
- Maximum Alignmentを用いて 492,993 文対を収集
  - 単語アライメントの閾値：単語間類似度が0.49以上
  - 文アライメントの閾値：文間類似度が0.53以上

	難解（English Wikipedia）	平易（Simple English Wikipedia）
0.9	Woody Bay Station was purchased by the Lynton …	Woody Bay Station was bought by the Lynton …
0.7	Miró has been a significant influence on late 20th-century art, in particular the American abstract expressionist artists such as Motherwell, … and others.	Miró was a significant influence on late 20th-century art, in particular the American abstract expressionist artists.
0.6	The couple has four children:	She has two daughters and two sons.

実験設定：外的評価

- 統計的機械翻訳の枠組みでのテキスト平易化

- トレーニング：テキスト平易化コーパス
- チューニング：MERT、無作為抽出した500文対
- テスト：BLEU、Hwangらの人手のラベル付きデータ
  - Good（両方向含意）277文対
  - Good Partial（片方向含意）281文対
- SMTツール：Moses（Phrase-based SMT）
- 言語モデル：KenLm (5-gram)、Simple English Wikipedia

BLEUを3.2ポイント改善

	文対数	平均文長		BLEU	
		難解	平易	G	G+GP
Baseline (None)				42.1	22.3
Zhu et al. (2010)	107,516	21.2	17.4	42.0	22.1
Coster and Kauchak (2011)	136,862	23.6	21.1	44.3	23.8
Hwang et al. (2015)	284,238	26.0	19.8	43.9	23.1
Ours	492,493	25.3	17.9	47.5	26.3

- 本研究で構築したコーパスで学習したモデルがBLEUを大きく改善
- 我々のコーパスは難解な文と平易な文の平均文長の差が大きい
  - English Wikipedia全体の平均文長：25.1
  - Simple English Wikipedia全体の平均文長：16.9
- Maximum Alignment は文長に関わらず適切に文間類似度を計算できる

提案手法はコーパスサイズに関係なくBLEUが高い

	文対数	平均文長		BLEU	
		難解	平易	G	G+GP
Zhu et al. (2010)	100,000	21.2	17.4	41.8	22.1
	107,516	21.2	17.4	42.0	22.1
Coster and Kauchak (2011)	100,000	23.7	21.1	43.8	23.4
	136,862	23.6	21.1	44.3	23.8
Hwang et al. (2015)	100,000	25.3	21.2	42.9	22.7
	200,000	25.6	20.5	43.1	22.7
	300,000	26.1	19.7	42.9	22.7
	391,116	26.5	19.4	43.1	22.8
Ours	100,000	23.9	21.8	43.2	23.6
	200,000	24.7	20.1	45.7	24.8
	300,000	25.2	19.0	46.4	25.3
	492,493	25.3	17.9	47.5	26.3

テキスト平易化の実例

Input	Mozart's Clarinet Concerto and Clarinet Quintet are both in A major, and generally Mozart was more likely to use clarinets in A major than in any other key besides E-flat major.
Reference	Mozart used clarinets in A major often.
Zhu et al.	Mozart's Clarinet Concerto and Clarinet Quintet are both in A major, and generally Mozart which he more likely to use clarinets in A major than in any other key besides E-flat major.
Coster and Kauchak	Mozart was Clarinet Concerto and Clarinet Quintet are both in A major, and Mozart used clarinets in A major often.
Hwang et al.	Mozart's Clarinet Concerto and Clarinet Quintet are both in A major, and generally Mozart was more likely to use clarinets in A major than in any other key besides E-flat major.
Ours	Mozart's Clarinet Concerto and Clarinet Quintet are both in A major, and Mozart used clarinets in A major often.

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

- 低コストで高精度なテキスト平易化コーパスを構築した。
  https://github.com/tmu-nlp/sscorpus

- 多対一の単語アライメントに基づく文間類似度計算手法の有効性を実験的に示した。
- 我々のコーパスで学習すると、既存のコーパスを用いる場合と比較して、統計的機械翻訳の枠組みでのテキスト平易化でBLEUを 3.2 改善できた。

- 今後は Simple English Wikipedia に頼らず、生コーパスのみからテキスト平易化コーパスを構築する手法に拡張し、多言語化したい。