Problems in Current Text Simplification Research: New Data Can Help

Wei Xu and Chris Callison-Burch and Courtney Napoles

TACL 2015

読み人：梶原智之（首都大小町研D1）

Problems in Current Text Simplification Research: New Data Can Help

- みんなSimple Wikipedia使ってますけど、これって本当に平易なテキストですか？
- 調べてみたら、そんなにシンプルではない。
- Text Simplification のための Simple English Wikipediaに頼らない良いコーパスを提案する。

https://newsela.com/data/

Text Simplification

- 複雑なテキストを書き換えて分かり易くする
  - 障がい者のための読解支援　(Inui et al. 2003)
  - ノンネイティブのための読解支援
  - 非専門家のための読解支援
- 各種NLPタスクに前処理をして性能を改善する
  - 自動要約
  - 情報抽出　(Miwa et al. 2010)
  - 機械翻訳

長くて複雑なテキストの書き換えは
人間のためにも機械のためにも有用

PWKP: Parallel Wikipedia Simplification Corpus

- 対訳コーパス（Wikipedia / Simple English Wikipedia）
- TS as a SMT

1. 記事単位のアライメント
   - 言語間リンク
   - 記事タイトル (Coster and Kauchak 2011, Kauchak 2013)
2. 文単位のアライメント
   - TF-IDF
   - 段落のアライメントを取り、順序制約付きでTF-IDF
   - 辞書をつかって単語の類似度まで見る (Hwang et al. 2015)

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. A Monolingual Tree-based Translation Model for Sentence Simplification. Proceedings of the 23rd International Conference on Computational Linguistics, pp. 1353–1361, 2010.

SEWは最善には及ばない

1. 文単位のアライメントにエラーが多い
2. 不十分な平易化が多い
3. 他の分野に十分一般化できない

- 自発的かつ共同で作られた百科事典
- “子どもや英語学習者のために”
  という以外の明確なガイドラインがない

PWKPの200文を人手でチェック

Not Aligned (17%)		[NORM] The soprano ranges are also written from middle C to A an octave higher, but sound one octave higher than written. [SIMP] The xylophone is usually played so that the music sounds an octave higher than written.
Not Simpler (33%)		[NORM] Chile is the longest north-south country in the world, and also claims of Antarctica as part of its territory. [SIMP] Chile, which claims a part of the Antarctic continent, is the longest country on earth.
		[NORM] Death On 1 October 1988, **Strauss** collapsed while hunting with the Prince of Thurn and Taxis in the Thurn and Taxis forests, east of Regensburg. [SIMP] Death On October 1, 1988, **Strauß** collapsed while hunting with the Prince of Thurn and Taxis in the Thurn and Taxis forests, east of Regensburg.
Real Simplification (50%)	**Deletion Only (21%)**	[NORM] This **article** is a list of the 50 U.S. states **and the District of Columbia** ordered by population density. [SIMP] This is a list of the 50 U.S. states**,** ordered by population density.
	Paraphrase Only (17%)	[NORM] In 2002, both Russia and China also had **prison populations in excess of 1 million.** [SIMP] In 2002, both Russia and China also had **over 1 million people in prison.**
	Deleltion + (12%) Paraphrase	[NORM] All adult Muslims**, with exceptions for the infirm, are required to offer** Salat prayers five times **daily**. [SIMP] All adult Muslims **should do** Salat prayers five times **a day**.

PWKPの200文を人手でチェック

Not Aligned (17%)		[NORM] The soprano ranges are also written from middle C to A an octave higher,
Not Simpler (33%)		of Thurn and Taxis in the Thurn and Taxis forests, east of Regensburg.
Real Simplification (50%)	**Deletion Only (21%)**	[NORM] This **article** is a list of the 50 U.S. states **and the District of Columbia**
	Paraphrase Only (17%)	
	Deleltion + Paraphrase (12%)	[SIMP] All adult Muslims **should do** Salat prayers five times **a day**.

- PWKPを学習したMosesは原文の22%を書き換えなかった（Wubben et al. 2012）
- コーパスの27%は原文と平易文が同じ（Coster and Kauchak 2011）

- ボランティアが基準もなくやっている
- 百科事典だから専門用語が多い

Newsela Corpus

- 専門の記者によって書かれた平易な記事
- ニュース記事を4レベルの子ども向けに書き換え
- アメリカの教育課程に則った英語レベル

Grade Level	Lexile Score	Text
12	1400L	**Slightly more** fourth-graders **nationwide are reading proficiently compared with** a decade ago**, but** only a third of them **are now reading well**, according to a new report.
7	1070L	Fourth-graders **in most states are better readers than they were a decade ago. But** only a third of them **actually are able to read well**, according to a new report.
6	930L	Fourth-graders **in most states** are better readers than they were a decade ago. But **only a third of them** actually **are able to** read well**, according to a new report**.
4	720L	**Most** fourth-graders are better readers than **they were** 10 years ago. But **few of them can actually** read well.
3	510L	Fourth-graders are better readers than 10 years ago. But few of them read well.

PWKP vs. Newsela

	PWKP	Newsela
記者	ボランティア	専門の記者
基準	子どもや外国人向け（なし）	アメリカの教育課程準拠（あり）
分野	百科事典	ニュース
コーパスタイプ	コンパラブルコーパス	パラレルコーパス

- 文単位のアライメントが確実に取れている
- ちゃんとした人がちゃんとした基準で書いてる
- 専門用語ばっかり出てくるわけではない

基本統計と語彙の比較

基本的な統計	Newsela					PWKP	
	Original	Simp-1	Simp-2	Simp-3	Simp-4	Normal	Simple
Total #sents	56,037	57,940	63,419	64,035	64,162	108,016	114,924
Total #tokens	1,301,767	1,126,148	1,052,915	903,417	764,103	2,645,771	2,175,240
Avg #sents per doc	49.59	51.27	56.12	56.67	56.78	—	—
Avg #words per doc	1,152.01	996.59	931.78	799.48	676.2	—	—
Avg #words per sent	**23.23**	19.44	16.6	14.11	**11.91**	***24.49**	***18.93**
Avg #chars per word	**4.32**	4.28	4.21	4.11	**4.02**	**5.06**	**4.89**

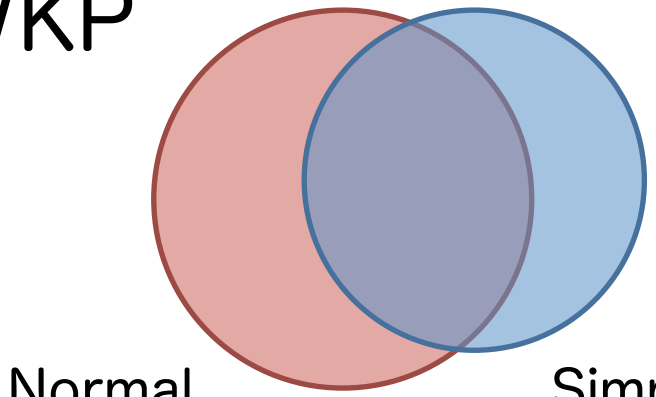
Newsela	Original	Simp-1	Simp-2	Simp-3	Simp-4
#words (avg. freq)	**39,046 (28.31)	33,272 (28.64)	29,569 (30.09)	24,468 (31.17)	20,432 (31.45)
Original	0	724 (1.19)	815 (1.25)	720 (1.32)	*583 (1.33)
Simp-1	6,498 (1.38)	0	618 (1.08)	604 (1.15)	521 (1.21)
Simp-2	10,292 (1.67)	4,321 (1.32)	0	536 (1.13)	475 (1.16)
Simp-3	15,298 (2.14)	9,408 (1.79)	5,637 (1.46)	0	533 (1.14)
Simp-4	**19,197 (2.60)	13,361 (2.24)	9,612 (1.87)	4,569 (1.40)	0

PWKP	Normal	Simple
#words (avg. freq)	95,111 (23.91)	78,009 (23.88)
Normal	0	6,669(1.31)
Simple	23,771 (1.42)	0

Simpleに現れる単語のうち
6,669語はNormalに現れない

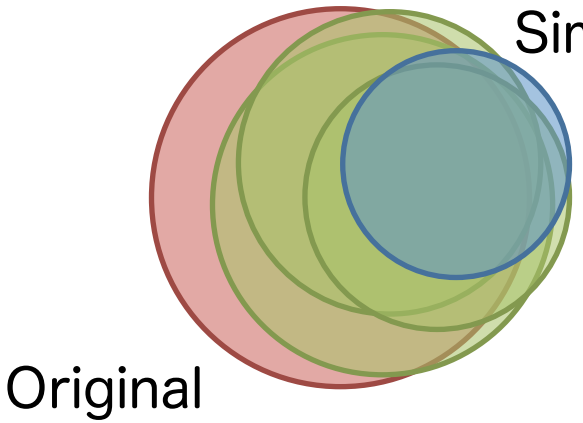
語彙の比較のイメージ

- PWKP



Simpleの語彙は
約半分をNormalの
語彙と共有している

- Newsela



よりSimpleな語彙は
よりComplexな語彙に
ほぼ包含されている

Log-odds-ratio analysis of words

Linguistic class	Newsela - Original	Wikipedia (PWKP) - Normal
Punctuation	, "– ; ' ( )	, ; –
Determiner/Pronoun	**which** we an such who i that a whose	**which** whom
Contraction	's	
Conjunction	**and while although**	**and although while**
Prepositions	**of as including with** according by among in despite	**as with** following to **of within** upon **including**
Adverb		currently approximately initially primarily subsequently typically thus formerly
Noun	percent director data research decades industry policy development state decade status university residents	film commune footballer pays-de-la-loire walloon links midfielder defender goalkeeper
Adjective	federal potential recent executive economic	northern northwestern southwestern external due numerous undated various prominent
Verb	advocates based access	referred derived established situated considered consists regarded having

Top 50 tokens associated with the complex text

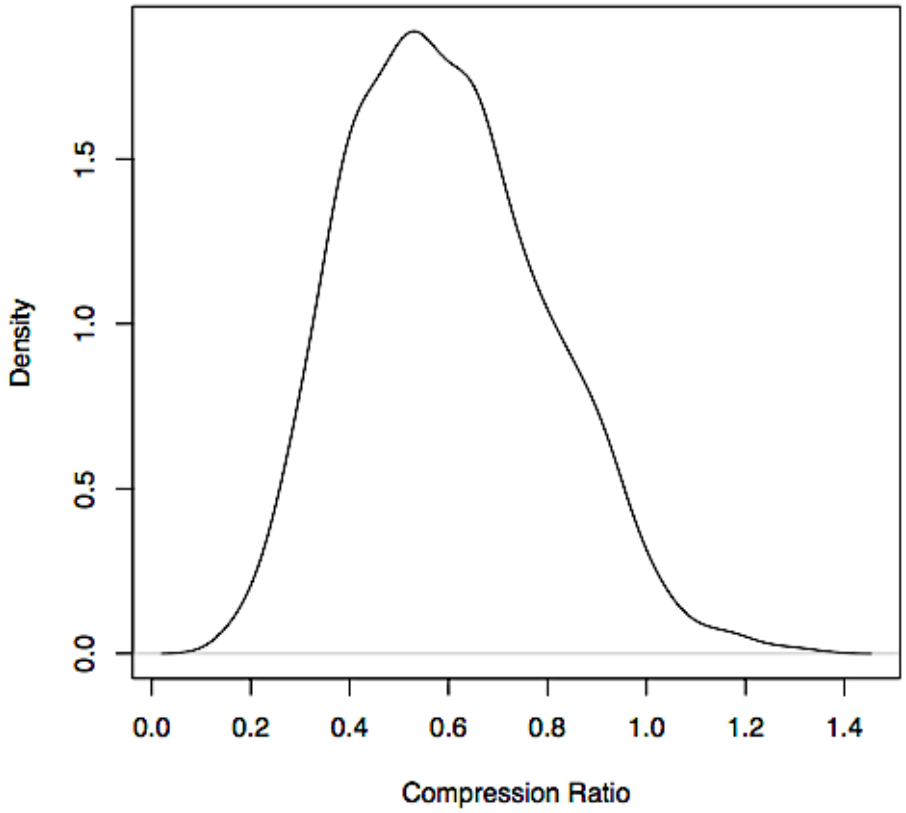
Log-odds-ratio analysis of words

Linguistic class	Newsela - Simp4	Wikipedia (PWKP) - Simple
Punctuation	.	.
Determiner/Pronoun	**they it he she** them **lot**	**it he they lot** this **she**
Conjunction		because
Adverb	also not **there** too **about very** now then how	**about very there**
Noun	**people** money scientists government things countries rules problems group	movie **people** northwest north region loire player websites southwest movies football things
Adjective	**many important big** new used	**big** biggest famous different **important many**
Verb	**is are can** will **make get** were wants **was called** help hurt be **made like** stop want works do live	found **is made called** started pays said **was** got **are like get can** means says has went comes **make** put used

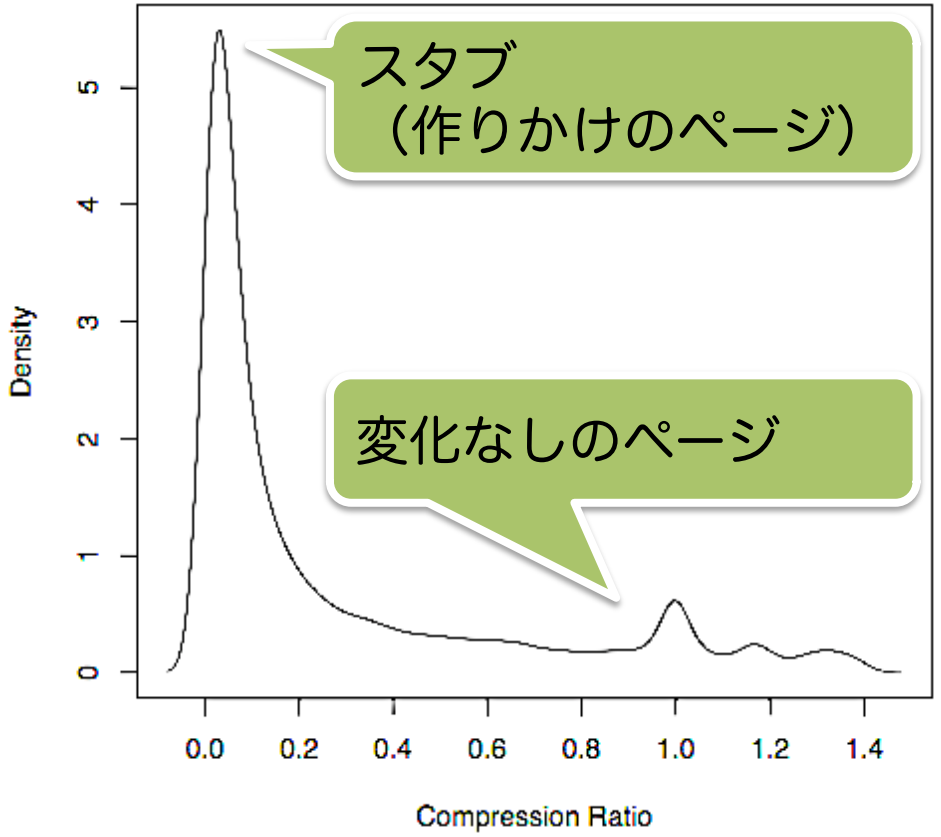
Top 50 tokens associated with the simplified text

Document-level Compression

Newsela



Wikipedia



Problems in Current Text Simplification Research: New Data Can Help

- Simple English Wikipedia に替わる Text Simplificationのためのコーパスを提案

  https://newsela.com/data/

- 専門家が教育課程の英語レベルを基準に書き換え
- パラレルコーパスなのでアライメントエラーなし
- 難解文の語彙が平易文の語彙を包含している