第7回NLP東京Dの会　http://cl.sd.tmu.ac.jp/~kajiwara/d7.pdf

# 語彙的換言を用いたテキスト平易化

首都大学東京 小町研
D2　梶原 智之

# Text Simplification

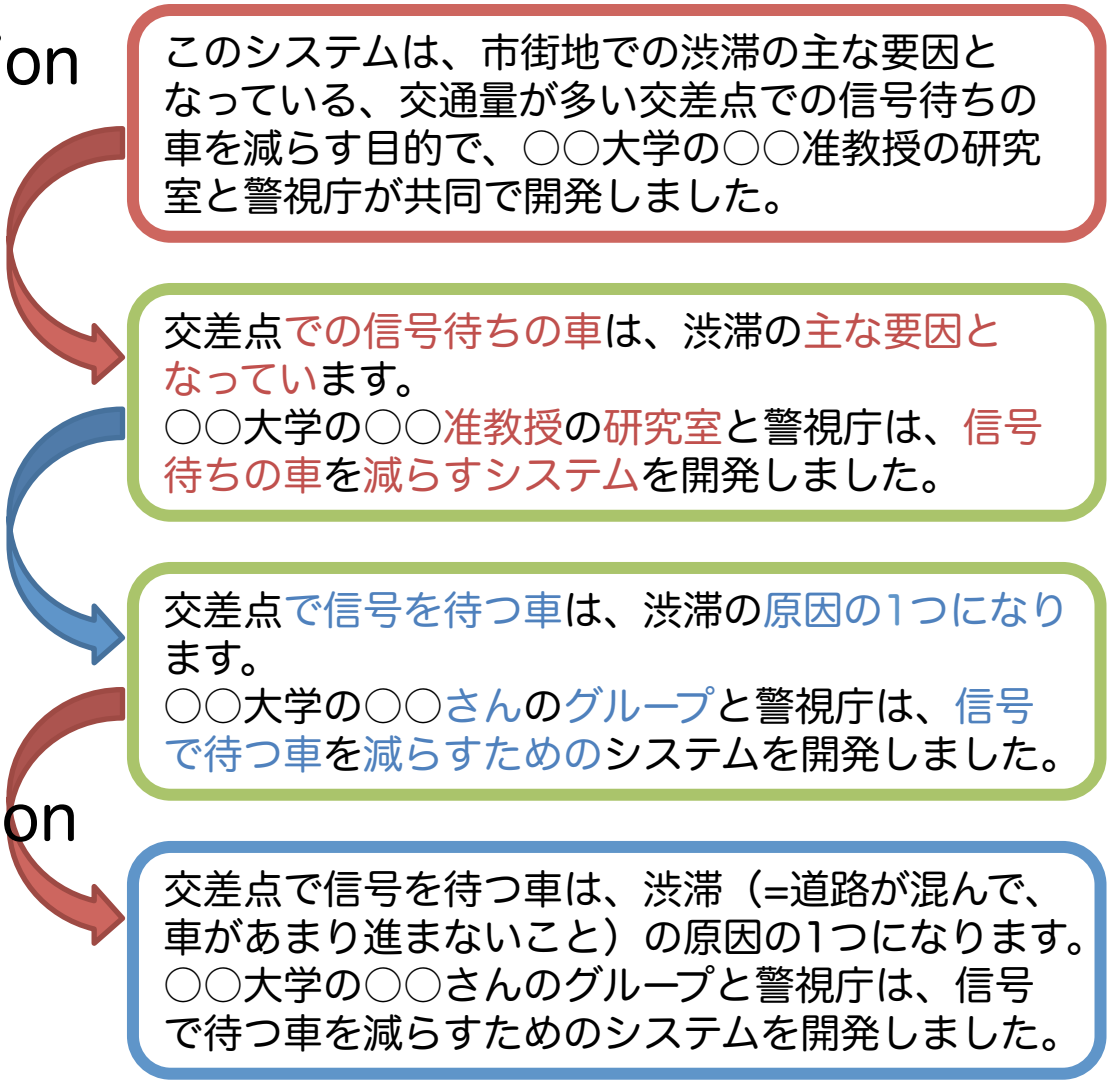
- English Wikipedia: Alfonso Perez
  - Alfonso Perez ~~Munoz, usually referred to as Alfonso,~~ is a former Spanish footballer~~, in the striker position~~.
- Simple English Wikipedia: Alfonso Perez
  - Alfonso Perez is a former Spanish football player.

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

# Text Simplification Pipeline

1. Syntactic Simplification
   - 文分割
   - 文圧縮
2. Lexical Simplification
   - 語句の言い換え
   - フレーズベースSMT
3. Explanation Generation
   - 辞書引き



# 語彙的換言を用いたテキスト平易化
# ~ Lexical Simplification ~

1. Lexical Approaches
2. PBSMT Approaches
3. Evaluation for Text Simplification
4. Resources for Text Simplification
5. 僕がやってきたこと

# 1. Lexical Approaches

入力：難解な語句
出力：平易な語句

| | | | |
|---|---|---|---|
| Devlin and Tait | 1998 | | The Use of a Psycholinguistic Database in the Simplification of Text for Aphasic Readers |
| De Belder et al. | 2010 | | Text Simplification for Children |
| Yatskar et al. | 2010 | NAACL | For the Sake of Simplicity: Unsupervised Extraction of Lexical Simplification form Wikipedia |
| Biran et al. | 2011 | ACL | Putting it Simply: a Context-Aware Approach to Lexical Simplification |
| Horn et al. | 2014 | ACL | Learning a Lexical Simplifier Using Wikipedia |
| Glavaš and Štajner | 2015 | ACL | Simplifying Lexical Simplification: Do We Need Simplified Corpora? |
| Paetzold and Specia | 2016 | AAAI | Unsupervised Lexical Simplification for Non-Native Speakers |
| Pavlick and Callison-Burch | 2016 | ACL | Simple PPDB: A Paraphrase Database for Simplification |

# 初期の語彙平易化

- Devlin and Tait 1998
- De Belder et al. 2010

Input Word →

1. 同義語辞書から言い換え候補を獲得
2. 候補の中から頻度の高い単語を選択

→ Simplified Word

非常に単純だが、これが語彙平易化の考え方の基本
どう「言い換えを集める」か、どう「ランキングする」か

# Simple English Wikipedia 出現以降の語彙平易化

| 手法 | | 候補の獲得 | ランキング |
|---|---|---|---|
| Yatskar et al. | 2010 | SEWの編集履歴 | |
| Biran et al. | 2011 | 分布類似度 | EWとSEWの出現頻度比、単語長 |
| Horn et al. | 2014 | パラレルコーパス（EW／SEW）＋単語アライメント | SVM-rank：単語アライメント確率、SEWの単語出現頻度、SEWのn-garm言語モデル |
| Glavaš and Štajner | 2015 | 単語分散表現のCOS類似度 | 平均ランキング：入力と候補のCOS、入力と文脈のCOSの平均、情報量=-log(freq(x)/freq(X))、n-gram言語モデル |
| Paetzold and Specia | 2016 | 単語分散表現のCOS類似度（同じ品詞のみ） | 字幕コーパスを用いた5-gram言語モデル |
| Pavlick and Callison-Burch | 2016 | PPDB（換言辞書） | 多クラスロジスティック回帰：EWとSEWの単語出現頻度比、単語長、音素数、換言確率、GoogleNgramの単語出現頻度、単語分散表現のCOS類似度 |

# 2. PBSMT Approaches

入力：難解な文
出力：平易な文

| English Approaches | | | |
|---|---|---|---|
| Coster and Kauchak | 2011 | ACL | Simple English Wikipedia: A New Text Simplification Task |
| Coster and Kauchak | 2011 | | Learning to Simplify Sentences Using Wikipedia |
| Wubben et al. | 2012 | ACL | Sentence Simplification by Monolingual Machine Translation |
| Štajner et al. | 2015 | ACL | A Deeper Exploration of the Standard PB-SMT Approach to Text Simplification and its Evaluation |
| Non-English Approaches（ポルトガル語、スペイン語、日本語） | | | |
| Specia | 2010 | | Translating from Complex to Simplified Sentences |
| Štajner et al. | 2015 | | Automatic Text Simplification for Spanish: Comparative Evaluation of Various Simplification Strategies |
| Goto et al. | 2015 | | Japanese News Simplification: Task Design, Data Set Construction, and Analysis of Simplified Text |

# 統計的機械翻訳の枠組みでのテキスト平易化

Coster and Kauchak 2011a

- パラレルコーパス + GIZA + Moses + BLEU
- English Wikipedia と Simple English Wikipedia のパラレルコーパスからGIZA++で平易化知識（難解なフレーズと平易なフレーズのペア）を獲得
- コーパスの一部をリファレンスとしてBLEUで評価

Coster and Kauchak 2011b

- 翻訳モデルの改良：フレーズの削除
- GIZA++による単語アライメントの際に、ある語句がNULLにアラインされたという情報を記憶

Wubben et al. 2012　（PBMT-R）

- リランキングの導入：たくさん書き換えられた文を出力
- Mosesの出力する10-bestの中から入力文との編集距離が最大の文を選択

## Štajner et al. 2015 a

PBSMTを用いたテキスト平易化における
パラレルコーパスの質と量が結果に与える影響を調査

| S-BLEU | Size of the training set | | | | |
|---|---|---|---|---|---|
| | 2,000 | 4,000 | 6,000 | 8,000 | 10,000 |
| [0,0.3] | 56.38 | 56.38 | 56.15 | 57.75 | 57.89 |
| (0.3,0.4] | 60.89 | 61.35 | 61.76 | 61.52 | 61.37 |
| (0.4,0.5] | 61.27 | 61.36 | 61.74 | 61.55 | 62.11 |
| (0.5,0.6] | 60.96 | 61.30 | 61.52 | 61.77 | 61.98 |
| (0.6,0.7] | 60.96 | 61.30 | 61.60 | 61.69 | 61.80 |
| (0.7,0.8] | 61.56 | 61.38 | 61.67 | 61.77 | 61.89 |
| (0.8,0.9] | 61.54 | 61.49 | 61.51 | 61.57 | 61.61 |
| (0.9,1] | 61.57 | 61.57 | 61.59 | 61.55 | 61.55 |

| System | G | M | S |
|---|---|---|---|
| Original | 4.85 | - | 2.60 |
| 0.3-02 | 4.03 | 3.95 | 2.57 |
| 0.3-10 | 4.20 | 4.03 | 2.85 |
| 0.6-02 | 4.50 | 4.45 | 2.68 |
| 0.6-10 | 4.43 | 4.48 | 2.72 |
| 1.0-02 | 3.25 | 2.92 | 2.45 |
| 1.0-10 | 2.92 | 2.95 | 2.53 |

- 類似度がとても低い文対を使うと良くない
- データの量は結果に大きな影響を与えない
- 中ぐらいの類似度の文対を上手く選ぶと良い

# 3. Evaluation for Text Simplification

| Lexical Approach | | | |
|---|---|---|---|
| Specia et al. | 2012 | SemEval | Task 1: English Lexical Simplification |
| De Belder and Moens | 2012 | CICLing | A Dataset for the Evaluation of Lexical Simplification |
| Horn et al. | 2014 | ACL | Learning a Lexical Simplifier Using Wikipedia |
| Paetzold and Specia | 2016 | LREC | Benchmarking Lexical Simplification Systems |

| PBSMT Approach | | | |
|---|---|---|---|
| Xu et al. | 2016 | TACL | Optimizing Statistical Machine Translation for Text Simplification |

# 初期の語彙平易化の評価用データ

- Specia et al. 2012
  - 201単語 × 10文脈 × 5人のアノテータ
  - 語彙的換言タスクの評価用データセットを並び替え
  - アノテータが入力単語と換言候補を難易度で並び替え
  - ５人の平均ランキングでデータセットを構築
- De Belder and Moens 2012
  - 43単語 × 10文脈 × 5人のアノテータ
  - 語彙的換言タスクのデータから難解な単語のみ抽出
  - クラウドソーシングでアノテータを採用
  - 各アノテータの信頼度を考慮しながら５人分のランキングを統合してデータセットを構築

# 語彙平易化の自動評価

- Horn et al. 2014
  - 500単語 × 1文脈 × 50人のアノテータ
  - EW/SEWのパラレルコーパスから単語を無作為抽出
  - アノテータが文脈も見つつ平易な言い換えを1語付与
- Paetzold and Specia 2016b　（BenchLS）
  - De Belder and MoensとHorn et al.のデータセットを組み合わせた929文　※ 7.37（平易語／難解語）

| BenchLSによる評価結果 | 手法の概要 | Precision | Accuracy | Changed |
|---|---|---|---|---|
| Biran et al. 2011 | 分布類似度, 頻度比 | 0.124 | 0.123 | 0.999 |
| Horn et al. 2014 | GIZA, SVM-rank | **0.546** | 0.341 | 0.795 |
| Glavaš and Štajner 2015 | COS, Avg-rank | 0.480 | 0.252 | 0.772 |
| Paetzold and Specia 2016 | COS, 5-gramLM | 0.416 | **0.416** | **1.000** |

# テキスト平易化の自動評価

- リーダビリティ: FRE, FKG
- 意味や文法: BLEU

| 人手評価との相関 (Spearman's $\rho$) | 文法 | 意味 | 難易度 |
|---|---|---|---|
| FKG | -0.002 | 0.136 | 0.147 |
| BLEU | **0.589** | **0.701** | 0.111 |
| SARI | 0.342 | 0.397 | **0.343** |

- Xu et al. 2016
  - クラウドソーシングで8人のマルチリファレンスを作成
  - 入力と出力とマルチリファレンスを比較する評価尺度
  - SARI = $(F_{add} + F_{keep} + P_{del}) / 3$

| テキスト平易化の評価 | 手法の概要 | FKG | BLEU | SARI |
|---|---|---|---|---|
| English Wikipedia | | 12.88 | **99.05** | 26.05 |
| Simple English Wikipedia | | 11.25 | 66.75 | 38.42 |
| Wubben et al. 2012 | Moses+リランキング | 11.10 | 63.12 | 33.77 |
| Xu et al. 2016 | PPDB+SARIチューニング | **10.90** | 72.36 | **37.91** |

# 4. Resources for Text Simplification

| | | | |
|---|---|---|---|
| ツール（語彙平易化パイプラインの実装） | | | |
| Paetzold and Specia | 2015 | ACL | LEXenstein:<br>A Framework for Lexical Simplification |
| 言い換え辞書（難解な語句と平易な語句のペア） | | | |
| Pavlick and Callison-Burch | 2016 | ACL | Simple PPDB:<br>A Paraphrase Database for Simplification |
| パラレルコーパス（難解な文と平易な文のペア） | | | |
| Zhu et al. | 2010 | COLING | A Monolingual Tree-based Translation Model for Sentence Simplification |
| Coster and Kauchack | 2011 | ACL | Simple English Wikipedia: A New Text Simplification Task |
| Hwang et al. | 2015 | NAACL | Aligning Sentences form Standard Wikipedia to Simple Wikipedia |
| Xu et al. | 2015 | TACL | Problems in Current Text Simplification Research: New Data Can Help |

# ツール & 言い換え辞書

Paetzold and Specia 2015　（LEXenstein）

- 語彙平易化パイプライン構築のためのフレームワーク
- ４つのタスクについて多くの手法が実装されている
  1. Complex Word Identification
  2. Substitution Generation
  3. Substitution Selection
  4. Substitution Ranking

https://github.com/ghpaetzold/LEXenstein

Pavlick and Callison-Burch 2016　（Simple PPDB）

- 450万フレーズペア
- 言い換え確率、難易度、難解な語句、平易な語句

http://www.seas.upenn.edu/~epavlick/data.html

# パラレルコーパス

- English Wikipedia と Simple English Wikipedia
  - Zhu et al. 2010　（6.5万記事→10万文対）

https://www.ukp.tu-darmstadt.de/data/sentence-simplification/simple-complex-sentence-pairs/

  - Coster and Kauchak 2011　（1万記事→13万文対）

http://www.cs.pomona.edu/~dkauchak/simplification/

  - Hwang et al. 2015　（2.2万記事→28万文対）

http://ssli.ee.washington.edu/tial/projects/simplification/

- Simple English Wikipedia を使わないもの
  - Xu et al. 2015　（Newselaコーパス）
  - ニュース記事を4段階の難易度に人手で書き換えたもの

https://newsela.com/data/

# 5. 僕がやってきたこと

- Lexical Approach
  - 語釈文を用いた小学生のための語彙平易化（2015, 情報処理学会論文誌）
  - Evaluation Dataset and System for Japanese Lexical Simplification（2015, ACL-SRW）
  - Controlled and Balanced Dataset for Japanese Lexical Simplification（2016, ACL-SRW）
- PBSMT Approach
  - Building a Monolingual Parallel Corpus for Text Simplification Using Sentence Similarity Based on Alignment between Word Embeddings（2016, COLING）
  - 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（2016, 12月のNL研）

# 日本語の語彙平易化

- 語釈文を用いた小学生のための語彙平易化
  - 小学国語辞典の語釈文は平易に書かれている
  - 「見出し語 → 語釈文」の平易化知識を獲得
  - LEXenstein [Paetzold+ 2015] にも一応実装されている

- 評価用データセット
  - 201単語 × 10文脈 × 5人　※ 4.30（平易語/難解語）
  - 日本語の均衡コーパス（BCCWJ）から構築

# 英語のテキスト平易化

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.

# 単語アライメントに基づく文間類似度

| 文間類似度計算手法 | 手法の概要 | 両方向含意 vs. 他 | | 片方向含意 vs. 他 | |
|---|---|---|---|---|---|
| | | MaxF1 | AUC | MacF1 | AUC |
| Zhu et al. 2010 | TFIDF+COS | 0.550 | 0.509 | 0.431 | 0.391 |
| Coster and Kauchak 2011 | TFIDF+COS +文の出現順序 | 0.564 | 0.495 | 0.415 | 0.387 |
| Hwang et al. 2015 | 辞書を使って 単語類似度を考慮 | 0.712 | 0.694 | 0.607 | 0.529 |
| Kajiwara and Komachi 2016 | 単語分散表現で 単語類似度を考慮 | <u>0.717</u> | <u>0.730</u> | <u>0.638</u> | <u>0.618</u> |

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



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

# https://github.com/tmu-nlp/sscorpus

- 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
- $\text{MaxSim}_{\text{sym}}$ 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 | 平易 |
|---|---|---|
| 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. |

# BLEUを3.2ポイント改善

| テキスト平易化コーパス | 文対数 | 平均文長 | | BLEU | |
|---|---|---|---|---|---|
| | | 難解 | 平易 | G | G+GP |
| Baseline (None) | - | (25.1) | (16.9) | 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** |

| | |
|---|---|
| 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. |
| Ours | Mozart's Clarinet Concerto and Clarinet Quintet are both in A major, and Mozart used clarinets in A major often. |