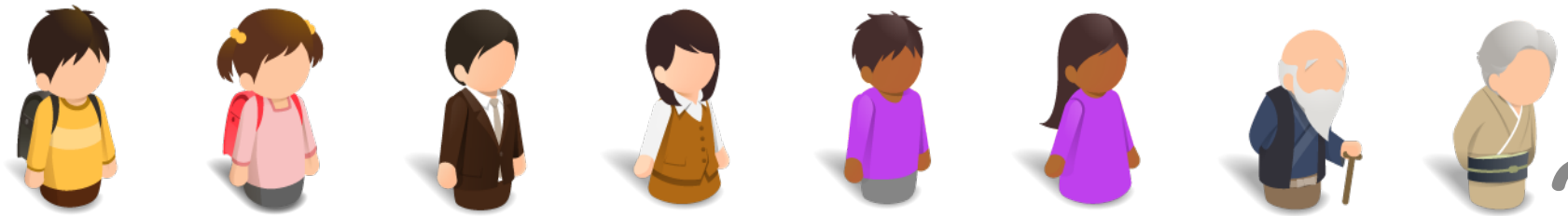
日本語の語彙平易化評価セットの構築

長岡技術科学大学
梶原 智之　山本 和英

研究の背景

アクセスは容易

理解は容易？



子供　大人　外国人　高齢者

研究の背景

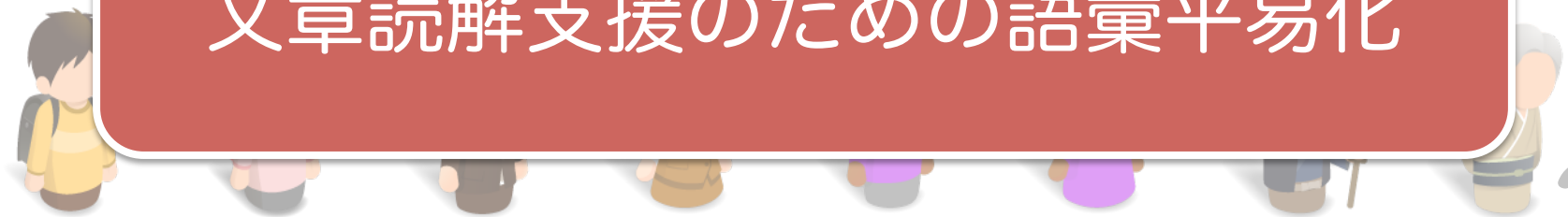
効率的な情報収集・知識獲得のため
言語能力の差を埋める技術が必要

大量・多様なテキストデータ

アクセスは容易

理解は容易？

文章読解支援のための語彙平易化



子供　大人　外国人　高齢者

語彙平易化

文中の難解な語をより平易な同義語に置換

四国に赴く

四国に行く

対象	評価尺度		赴く	行く
大人	単語親密度DB	難：1 → 易：7	5.0	6.469
子ども	学習基本語彙	難：– → 易：✓	–	✓
外国人	日本語能力試験	難：1 → 易：4	1	4
外国人	日本語教育語彙表	難：6 → 易：1	5	1

幅広い読者の文章読解を支援する

関連研究

- SemEval-2012: English Lexical Simplification Task [1]
  - 語彙平易化の評価のためのデータセットが構築される
  - 9つのシステムが参加：文脈を考慮して高頻度語に置換
- 公開されている英語の語彙平易化システム
  - Automatic Sentence Simplification Using Wikipedia [2]
  - Rewordify.com　(https://rewordify.com/)
- 公開されている英語の語彙平易化のための評価セット
  - Speciaらのデータセット [1]
  - De Belderらのデータセット [3]

[1] Lucia Specia et al. (2012) “Semeval-2012 Task 1: English Lexical Simplification” http://www.cs.york.ac.uk/semeval-2012/task1/
[2] Kristian Woodsend and Mirella Lapata (2011) “WikiSimple: Automatic Simplification of Wikipedia Articles” http://homepages.inf.ed.ac.uk/kwoodsen/demos/simplify.html
[3] Jan De Belder and Marie-Francine Moens (2012) “A Dataset for the Evaluation of Lexical Simplification” http://people.cs.kuleuven.be/~jan.debelder/lseval.zip

研究資源の公開の重要性

- 語彙平易化システムの公開 [4]
  - 読解支援を必要とする読者に語彙平易化の技術を届ける
  - http://www.jnlp.org/SNOW/S3

- 語彙平易化の評価のためのデータセットの公開
  - 従来の人手評価のコストと再現性の課題を解決し、適合率および再現率の自動評価の枠組みを提供する
  - 複数の語彙平易化システムの性能を直接比較する
  - http://www.jnlp.org/SNOW/E4

[4] 梶原智之, 山本和英 (2015) “日本語の語彙平易化システムの構築”

英語の語彙平易化評価セット

- SemEval-2007 English Lexical Substitution Task [5] の語彙的換言の評価のためのデータセットを並び替え
  - 対象語201語 × 10文脈 = 2,010文
  - アノテーション：5人の英語母語話者が文脈中で対象語の語彙的換言を列挙
  - <context>During the siege, George Robertson had appointed Shuja-ul-Mulk, who was a <head>bright</head> boy only 12 years old and the youngest surviving son of Aman-ul-Mulk, as the ruler of Chitral.</context>
  - Gold-Standard: intelligent 3; clever 3; smart 1;

[5] Diana McCarthy and Robert Navigli (2007) “SemEval-2007 Task 10: English Lexical Substitution Task”

英語の語彙平易化評価セット

- Speciaらの語彙平易化評価セット
  - 対象語201語 × 10文脈 = 2,010文
  - アノテーション：4-5人の非英語母語話者が文脈中で対象語の換言を難易度で並び替え
  - アノテーションの統合：各難易度ランクの平均値

1. {clear} {light} {bright} {luminous} {well-lit}
2. {well-lit} {clear} {light} {bright} {luminous}
3. {clear} {bright} {light} {luminous} {well-lit}
4. {bright} {well-lit} {luminous} {clear} {light}

Gold: {clear} {bright} {light, well-lit} {luminous}

e.g. AverageRank(clear) = (1+2+1+4) / 4 = 2
AverageRank(bright) = (3+4+2+1) / 4 = 2.5

英語の語彙平易化評価セット

- De Belderらの語彙平易化評価セット
  - 対象語43語 × 10文脈 = 430文
  - 十分に平易な対象語を削除
    （Simple English Wikipediaの基本語彙など）
  - アノテーション：5人の英語母語話者が文脈中で
    対象語の換言を難易度で並び替え
  - アノテータの募集：Amazon Mechanical Turk

日本語の語彙平易化評価セット

1．語彙的換言
データセットの構築 ⇒ 2．語彙平易化
データセットへの変換

- 語彙的換言データセットの構築：対象語の選定
  1. IPA辞書 ∩ JUMAN辞書の内容語（名詞、動詞、形容詞、副詞）
  2. 平易な語を削除
     ※学習基本語彙（小学生のための理解語彙）に含まれる語を削除
  3. 換言が存在しない語を削除
     ※内容語換言辞書（SNOW D2）に含まれない語を削除
  4. 低頻度語を削除
     ※新聞記事15年分での出現頻度が10未満の語を削除

対象語：名詞・動詞 75語、形容詞・副詞 50語（無作為抽出）

日本語の語彙平易化評価セット

- 語彙的換言データセットの構築：換言の付与
  - 各対象語に10種類の文脈を新聞記事から無作為に付与
  - 5人のアノテータが文脈中で対象語の言い換えを列挙
  - アノテータ：クラウドソーシングで募集（http://www.lancers.jp/）
  - 平均 5.38 語の語彙的換言が付与された（一致率：17.8%）

- 語彙的換言データセットの構築：付与された換言の評価
  - 5人のアノテータのうち3人以上が「適切な言い換えである」と回答した表現のみ採用
  - アノテータ：新たにクラウドソーシングで募集
  - 平均 4.50 語の語彙的換言が採用された（一致率：66.4%）

日本語の語彙平易化評価セット

- 語彙平易化データセットへの変換：難易度で並び替え
  - 5人のアノテータが文脈中で対象語とその換言を平易な順に並び替え（一致率：33.2%）
  - アノテータ：換言の評価の際に募った作業者
- 語彙平易化データセットへの変換：難易度ランクの統合
  - 5人の難易度ランクの平均値
- クラウドソーシング：のべ500人が作業
  - 換言の付与：のべ250人
  - 換言の評価と並び替え：のべ250人

日本語の語彙平易化評価セット

- 語彙的換言と語彙平易化の評価のためのデータセット
  二つの位置がピッタリ合ったところを【検出する】か、
  差を【検出する】かという部分だけが異なる。
  - 【検出する】発見する 1；検知する 4；見つける 1；
  - 平易 ← (見つける) (発見する・【検出する】) (検知する) → 難解

データセット	総文数	名詞	動詞	形容詞	副詞
Speciaら	2,010	580 (28.9%)	520 (25.9%)	560 (27.9%)	350 (17.4%)
De Belderら	430	100 (23.3%)	60 (14.0%)	160 (37.2%)	110 (25.6%)
梶原ら (SNOW E4)	2,330	630 (27.0%)	720 (30.9%)	500 (21.5%)	480 (20.6%)

データセットの特性

データセットの文脈依存性	
①：対象語が同じ文脈の組	10,485
②：①のうち換言リストが等しい組	1,593
③：②のうち難易度ランクが違う組	948
④：③のうち最も平易な語が違う組	463

15.2%

59.5%

48.8%

対象語と文脈	換言リスト（上段）と難易度ランク（下段）
グルメというのが、食のバブルであるとするなら、それは【とっくに】終わった文化である	すでに；既に；とうに；随分前に；前に；もう；
	{とっくに}{すでに}{もう}{既に}{とうに}{前に}{随分前に}
どうやら職場での飲酒は【とっくに】ばれていたらしい	とうの昔に；すでに；既に；とうに；随分前に；もう；
	{とっくに}{すでに}{既に もう}{とうの昔に}{とうに}{随分前に}
【とっくに】気付いているかもしれないが、写真中央にいるのはＦ１でもおなじみのナイジェル・マンセルだ	とうの昔に；すでに；既に；とうに；随分前に；もう；
	{すでに}{もう}{とっくに}{既に}{とうに}{とうの昔に 随分前に}

データセットの特性

データセットの文脈依存性	
①：対象語が同じ文脈の組	10,485
②：①のうち換言リストが等しい組	1,593
③：②のうち難易度ランクが違う組	948
④：③のうち最も平易な語が違う組	463

15.2%

59.5%

48.8%

データセットの特性	
換言リストの長さの平均	5.50
難易度ランクの平均	4.94
対象語よりも平易な語が存在する割合	69.4%
対象語と同じ難易度の語が存在する割合	18.0%
対象語よりも難解な語が存在する割合	83.5%

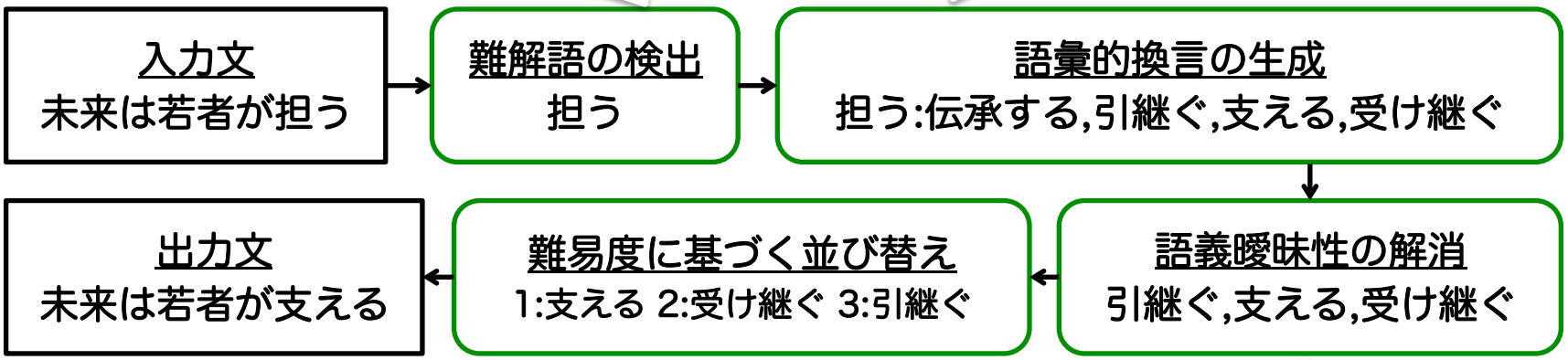
日本語の語彙平易化システムの構築

http://www.jnlp.org/SNOW/S3

Precision	Recall	F-measure
84.4 %	71.4 %	77.3 %

語彙的換言知識：
- 基本的意味関係の事例ベース
- 内容語換言辞書（SNOW D2）
- 動詞含意関係DB
- 日本語WordNet同義語DB

形態素解析：MeCab
平易語：学習基本語彙



難易度：単語親密度DB

述語項構造解析：SynCha
格フレーム辞書：京都大学格フレーム

日本語の語彙平易化評価セットの構築

http://www.jnlp.org/SNOW/E4

1. 語彙的換言データセットの構築

1. 対象語の選定
2. クラウドソーシングを用いた語彙的換言の列挙
3. 複数の作業者による作業結果の統合
   （クラウドソーシングを用いた語彙的換言の評価）

2. 語彙平易化データセットへの変換

1. クラウドソーシングを用いた平易化候補の難易度による並び替え
2. 複数の作業者による作業結果の統合

- 語彙的換言と語彙平易化の評価のためのデータセット

  二つの位置がピッタリ合ったところを【検出する】か、差を【検出する】かという部分だけが異なる。

  - 【検出する】発見する 1；検知する 4；見つける 1；
  - 平易 ← (見つける) (発見する・【検出する】) (検知する) → 難解