言い換え認識技術の評価に適した言い換えコーパスの構築指針

藤田 篤　　柴田 知秀　松吉 俊　　渡邉 陽太郎　　梶原 智之
情報通信研究機構　京都大学　山梨大学　日本電気株式会社　長岡技術科学大学

1　はじめに

人間の言語には，次の例に示すように，概ね同じ意味内容を表す異なる言語表現が多数存在する．このような関係にある表現を**言い換え** (paraphrase) と言う．

(1) a. 重傷を負う恐れがある
　　b. 大ケガをしてしまうかもしれない

(2) a. 料理がすぐに出てくる店に来ている
　　b. このお店ではあまり待たずに食べられる

計算機による言い換え処理の必要性は，情報検索，機械翻訳，文書要約等の自然言語応用技術における課題として古くから認知されていたが，1990 年代後半になり，それらの応用技術とは独立した要素技術としての研究が盛んに行われるようになってきた．それ以降，計算言語学・自然言語処理分野において，基礎的な分析から応用まで，また具体的なアプローチも言語学的なものから工学的なものまで多岐に渡る研究が行われてきた．これらは例えば次のように大別できる．

(a) 様々な言い換え現象の網羅・類型化
(b) 特定の種類の言い換えに関する事例研究
(c) 言い換え処理に必要な言語資源の開発
(d) 言い換え認識技術の開発
(e) 言い換え生成技術の開発
(f) 自然言語処理応用技術への適用・導入

我々は現在，日本語を対象として，上記 (d) の言い換え認識タスクにおける既存の手法の到達状況，および今後取り組むべき課題を明らかにしようとしている．これまでに，英語における同タスクの研究動向および NTCIR RITE-2 において構築された日本語におけるテキスト間含意関係認識技術の評価データ [61] の分析をふまえ，上記の分析を可能にする評価データが満たすべき要件，およびそれを担保するためのデータの構築方法について検討してきた．本稿では，それらを通じて得られた知見について報告する．

2　言い換えの範疇，関連する関係との違い

2.1　言語表現の意味

本稿の冒頭では，言い換えが「同じ意味内容を表す異なる言語表現」と述べた．ここでいう意味とは何か? 乾らによる解説論文 [31] は，言語表現の意味を次の 4 つのレベルに分けて説明している．

(a) 真理値意味論的意味/指示的意味 (denotation)
(b) 言外の意味/暗示的意味 (connotation)
(c) 参照対象/Saussure のレフェラン (référent)
(d) 語用論的効果/発語内行為 (illocutionary act)

各々の詳細な説明は同解説論文に譲るが，本稿では，上記の指示的意味の同一性をもって言い換えを定義する．これは，Saussure の用語を用いれば,「同じシニフィエ (signifié) を指す異なるシニフィアン (signifiant)」に対応する．すなわち，Frege が例に挙げた「明けの明星」と「宵の明星」はいずれも同じ「金星」を指示する (レフェランが等しい) が，異なる特殊な状況に関する意味を内包している (シニフィエが異なる) ため，言い換えではないとする．

ただし，言い換えの応用技術によっては，シニフィエのずれを許容して「意味が同じ」と見なせる場合もあるかもしれない．例えば，上の 2 つの表現を含む次の 2 文を考えよう．

(3) a. 明けの明星とも呼ばれる金星は太陽系で 2 番目に太陽に近い惑星である．
　　b. 宵の明星とも呼ばれる金星は太陽系で 2 番目に太陽に近い惑星である．

主節では同じ情報が同じように述べられており，シニフィエも等しい．違いは，従属節において，異なる別名 (上述のようにシニフィエも異なる) が例示されている点である．主節に据えられた「主たるシニフィエ」の同一性のみに基づいてこれらの 2 文を言い換えとみなして構わないような応用も存在すると思われる．

2.2　含意関係，推論

乾らによる解説論文 [31] の発表よりも後に，PASCAL におけるプロジェクト [13] を端緒として，**テキスト間含意関係** (textual entailment) に関する研究が活発に行われるようになった．これは，例えば次の 2 つの文の間の関係を取り扱う．

(4) t_1. 川端康成は「雪国」などの作品でノーベル文学賞を受賞した。
　　t_2. 川端康成は「雪国」の著者である。

この例のように，t_1 が真である場合に t_2 も常に真である，という関係が成り立つ場合に，t_1 は t_2 を含意し

ている[1]と言う．典型的な下位区分として，下位概念と上位概念の間の関係 (e.g.,「熟睡する」と「眠る」) や事態表現とその前提要件となる事態表現の関係 (e.g.,「忘れた」と「知っていた」) が挙げられる．また，t_1 と t_2 が互いに他方を含意する場合，すなわち両方向に含意関係が成立する場合は，t_1 と t_2 は言い換えである．

PASCAL で実施された RTE プロジェクトや，日本語を対象とした同種の NTCIR RITE-2 では，上に挙げたような，言語に関する知識および一般常識に照らして判断できるような事象のみでなく，次に示すような例も対象に含めている．

(5) t_1．うちの子はトマトをよく食べる。
t_2．うちの子はトマトが好きだ。

例 (4) とは異なり，この例については，t_2 が成立しない条件も考えられる．例えば，t_1 の理由として「親がよく食事に出し，残さないように躾けている」ということも考えうる．このような，人間ならば**推論** (infer) できるある程度蓋然性が高い関係 (inference) も，高度な自然言語処理の実現に欠かせない技術である [9].

3 各サブタスクの研究動向と課題

1 節で大別した言い換え研究のうち，言い換えそのものに関する工学的な処理・タスクに関するものは，(c) 言い換え処理に必要な言語資源の構築，(d) 言い換え認識，(e) 言い換え生成の 3 つである．本節では，各サブタスクに関する評価の研究動向と課題について述べる．

3.1 言い換え認識

言い換え認識 (paraphrase recognition/identification) タスクは，次のように定式化できる (図 1 も参照のこと).

入力: 同一言語の複数 (一般的に 2 つ) の異なる言語表現
出力: 入力された言語表現が同義であるか否か[2]

大規模なテキストデータ中の異なる言語表現で記述されている同一の意味内容を同定し，そのような多様性を吸収する技術は，情報検索，質問応答，複数文書要約などに有用である．

言い換え認識技術の開発・評価のために，上記の定式化のもとでの正負のラベル付きのテキスト対 (以下，事例) を収集し，**言い換えコーパス**を構築する研究が行われてきた [15, 62, 11, 63]．英語については，Microsoft

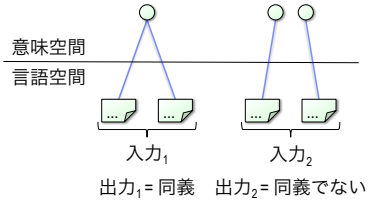

図 1: 言い換え認識タスクの模式図.

Research Paraphrase Corpus (MSRP) [15] が標準的な開発・評価データとして認識されている．PASCAL および TAC で開発されたテキスト間含意関係認識に関するデータ[3]にも，「互いに含意関係にあるテキスト対」，すなわち言い換え事例が含まれている．一方，日本語については，言い換え認識技術の開発・評価を目的として構築されたデータ[4]は，我々が知る限り存在しない．ただし，テキスト間含意関係認識技術の開発・評価 [48, 45, 58, 61, 47]，語彙の簡単化の評価 [32]，Web 上の言論の整理・可視化 [44] のために構築された各データの中に，言い換え事例も含まれている．これらの中では，NTCIR RITE-2 で構築されたデータ [61][5] が最も多くの研究グループに用いられている．

上記のような問題の定式化と，それに従う開発・評価データの公開により，主として機械学習に基づく言い換え認識手法の開発が促進された．本稿執筆時点で報告されている最高性能 (同義表現対の再現率と精度の F 値) は，MSRP については 84.1，RITE-2 (の “B”) については 69.3 である．MSRP に対する数値が非常に高い[6]のは，評価データに含まれる正例が負例よりも顕著に多い (66.5%) ためである，全事例に対して「同義」と判定する単純な手法 (以下，Eq 法) でも F 値約 80 を達成できてしまう．RITE-2 の評価データについては，548 事例中言い換え事例は 70 件であり，上述の Eq 法による F 値約 23 に比べれば，現時点の最高性能の値は比較的高い．まだ伸びしろはあるものの，このデータが元々，一方向の含意関係 (“F”)，両方向の含意関係 (“B”)，矛盾 (“C”)，無関係 (“I”) の 4 値分類を指向して作成されたこと，NTCIR RITE-2 では 4 クラスの F 値のマクロ平均が評価関数であったことを考慮すると，言い換え (“B”) のみに焦点を当てた開発により，ある程度は性能を改善できる可能性がある．

ただし，MSRP あるいは RITE-2 のデータに対する性能に基づいて各種手法を比較すると，高度な言語資源

[1] Logical entailment や，語用論における会話の含意 (implicature) とは異なる．

[2] 本稿で述べるのはもっぱら 2 値分類，すなわち「同義 (≡)」か「同義でない (≠)」かを特定するタスクである．ただし，近年，類似度を [0, 1] に定量化する Semantic Textual Similarity タスク [1] や，与えられた文脈において置換可能な同義語を複数の候補の中から選択する Lexical Substitution タスク [41] も提案されている．

[3] http://www.nist.gov/tac/data/

[4] 1 節 (a) の成果としての事例集 [20, 59] や 1 節 (b) の事例研究のための事例集 [18, 21] も構築されているが，言い換え認識技術の評価に必要な要件 (4.2 節) を満たすものではない．

[5] http://warehouse.ntcir.nii.ac.jp/openaccess/rite/10RITE-Japanese-wiki.html

[6] ACL Wiki: Paraphrase Identification

表 1: MSRP の評価データにおけるトークンの重複率.

群	事例数	重複率
同義	1147	0.715
同義でない	578	0.600

表 2: RITE-2 の評価データにおけるトークンの重複率.

群	事例数	重複率	
		t_1	t_2
B	70	0.726	0.712
F	205	0.397	0.733
C	61	0.449	0.589
I	212	0.405	0.489

や解析技術を用いる手法が，表層的な手がかりのみに基づく手法に対して顕著に高い性能を達成できておらず，評価データから何らかの影響を受けている可能性が疑われる．例えば，各事例におけるトークンの重複率を考えてみよう．MSRP については語[7]，RITE-2 のデータについては形態素[8]の表層形に基づいて，各事例における共通するトークンの数とテキスト中のトークン数の比 (重複率) を計算し，事例群ごとにマクロ平均を求めた結果を表 1 および表 2 に示す．この表が示す群間の顕著な差は，表層的な手がかりのみで問題がある程度解けるということを示唆している．

MSRP にはさらに深刻な問題が 2 つある．1 つ目は訓練データおよび評価データを構築する際に，編集距離が一定の範囲 (8〜20) のテキスト対のみを事例の候補としている点である．この制約はカバーできる言い換えの種類を極端に限定してしまう [15] ため，各手法で構築されたモデルも評価結果も，そのような制約のない言い換え現象に関する議論に耐えられない．2 つ目の問題は，正解ラベルが信頼性に欠けることである．言い換えとしての適否の判定を，厳密なガイドラインなしに作業者の直感のみに基づいて行ったため，評価データには，次の例のような，明らかに内容が異なるにもかかわらず「同義」と判定されている例も存在する [17].

(6) a. The stock rose $2.11, or about 11 percent, to close Friday at $21.51 on the New York Stock Exchange.

b. PG&E Corp. shares jumped $1.63 or 8 percent to $21.03 on the New York Stock Exchange on Friday.

RITE-2 の評価データの全 548 事例のトークンの重複率を図 2 に示す．言い換え事例 (“B”) のトークンの重複率は，他の 3 群よりも顕著に高い．作業者がキーワード検索によって得た Wikipedia 文書から事例の候補を

[7]Mosesdecoder (2.1.1) の tokenizer.perl を使用.
[8]MeCab (0.996) を使用.

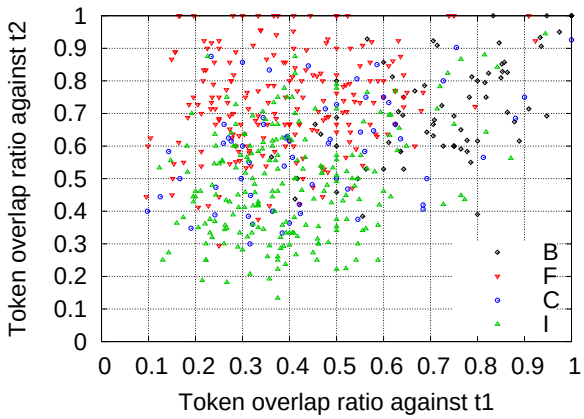

図 2: RITE-2 の評価データにおける各事例のトークンの重複率.

収集する際に，語の重複を手がかりにしてテキストを選択してしまった影響 (標本選択バイアス) だと思われる．その後の事例の選定および関係の判定は複数の人間によって独立に行われており，MSRP における明示的な制約ほどは現象の多様性を損なっていないと考えられるものの，識別は容易になってしまっている．一方向の含意関係の対 (“F”) については，複数の節からなる t_1 の一部を切り取って t_2 としている事例が多いため，t_1 に対する重複率が低く，t_2 に対する重複率が高い傾向がある．矛盾の事例 (“C”) は，他の群とは異なり，人手で作例されたものである．多様な例を作例するように教示はされていたが，“B” や “F” と区別しづらくするような工夫はされていない．

RITE-2 のデータについて特筆すべきは，一部分についてではあるが，個々の事例をよりプリミティブな[9]含意関係の連鎖に分解し，解決済/未解決の問題をより精密に分析することを可能にしている点である [33].

3.2 言い換え生成

言い換え生成 (paraphrase generation) タスクは，次のように定式化できる (図 3 も参照のこと).

入力: ある言語表現，目的に応じた評価基準

出力: 入力された言語表現と同じ言語の異なる言語表現の集合．ただし，入力された言語表現における基準を満たさない箇所を，意味を変えずに基準を満たすようにしたもの.

意味を保持したまま目的にあった多様な言語表現を生み出す技術は，テキストの簡単化，自動要約のサブタスクである文圧縮，機械翻訳や音声合成などの下流タスクの性能向上のための前処理などに有用である.

[9]「原始的な」あるいは「幼稚な」という意味ではなく,「原子的な (atomic)」という意味で用いる.

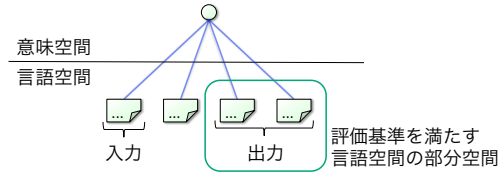

図 3: 言い換え生成タスクの模式図.

乾ら [31] は，上記の言い換え生成の問題は次の 2 つの部分問題 (処理) に分解できると述べている．

言い換え候補生成: 与えられた言語表現に対して，言語的に適格な種々の言い換えを網羅的に生成する処理

テキスト評価: 与えられた評価基準に基づいてテキストを評価する処理

しかし，目的に応じた制約を取り払い,「網羅的に」と言ったとたんに，言い換え候補生成の問題は，解くことも評価することも極端に困難になってしまう．

多様な言い換え候補を頑健かつ正確に生成するには，同義表現に関する膨大な知識 (3.3 節) を明示的にシステムに組み込む必要がある[10]．ただし，言い換え候補生成の能力は，言語知識の量だけでなく，言語知識の表現方法および適用方法 [42, 16, 57, 51, 66, 24] にも強く依存する．したがって，言い換え生成の方法論を論じる際は,「網羅的な」言い換えの生成を可能にするような知識の実現可能性についても言及する必要がある．

言い換え認識と同様に，出力すべき言い換えの集合を作成すれば網羅性を評価することができる，という考え方がある．しかし，何をもって，所与の言語表現に対する言い換えを「網羅した」と言えるかは自明ではない．この難しさゆえに，我々が知る限り，実際にこのようなアプローチで評価が行われた例は存在しない．例えば，文献 [19] では，特定の目的を想定しない言い換え候補生成の評価を行っているが，生成された言い換え候補における誤りの種類や分布に関する評価に留まっており，網羅性やその前提となる知識や生成手法への依存性に関する議論はなされていない．

特定の目的とそれに応じた評価基準を想定する場合は，手法の開発・評価ともに多少は現実的にできる．例えば，文圧縮 [12, 24] やテキスト簡単化 [66, 26] という目的を想定すると，入力に対して 1 つないし少数の参照出力を与えることができる．そして，そのようなデータを用いて，従来の機械学習法，例えば統計的機械翻訳におけるそれと同じ方法論で開発・評価を行うことができる．目的を達成することのみを考えればこのアプローチは合理的であるが，より適切な文圧縮，より適切な簡単化の出力が存在するとしてもそれを追求しない，という選択をしているに過ぎない．批判的に言えば，網羅性という課題からは目を背けている．

3.3 言い換え処理に必要な言語資源の開発

文献 [20] において，言い換えとみなせる種々の現象の類型 (1 節の (a))，およびそれらを実現するために必要な知識のうち，個々の語の素性や語間の関係に関する知識 (語彙知識) の類型が示されている．頑健かつ正確な言い換え処理 (特に言い換え生成) の実現には，これらの他にも，文法性を評価するための知識や事態間の関係知識など多様な知識が必要であるが，ここでは特に，**言い換え知識獲得** (paraphrase acquisition) に関する研究について概観する．

言い換え知識獲得タスクは，次のように定式化できる．

入力: コーパスや辞書等の様々な言語資源

出力: 同義表現集合 (対であることが多い) の集合

文あるいはそれよりも大きなテキストを出力の単位とする場合は，便宜的に言い換えコーパスの構築 (3.1 節) と呼ぶ[11]ことにし，以下では，次の例のような，文よりも小さなテキスト断片[12]を出力の単位と想定する．

(7) a. 重傷 ⇔ 大ケガ
b. 重傷を負う ⇔ 大ケガをする
c. 料理がすぐに出てくる
⇔ あまり待たずに食べられる

言い換え知識獲得は，言い換え表現対の候補の収集と，各々に対する検査の 2 段階の処理からなる．すなわち，言い換え認識を内包している．以下では，自動獲得できる言い換え知識の多様性の観点から，入力となる言語資源をコーパスに限定し，その種類ごとに，言い換え候補表現対の収集および同義性の検査に用いられてきた手がかりを紹介する．

単言語コーパス: カバレージの面で最も有望な知識源である．既存研究のほとんどが，分布仮説 [28] に基づいて，使用される文脈が類似する (分布類似度または文脈類似度が高い) 表現の対を言い換え表現対として獲得している．文脈要素としては，表層表現の左右のトークン n グラム [49, 6, 39]，構文

[10] 言い換え認識については，深層学習により，明示的な知識を持たずにモデルを学習できることが示されている [55].

[11] これらを “sentential paraphrase” という呼ぶ場合もあるが，言い換えの類型の観点を考慮すると，文の単位で初めて同義性が成り立つ場合でない限り，この呼び方は避けるべきである．例えば，例 (1) の文対は，“paraphrase sentences including lexical and phrasal paraphrases” という方が適切であろう．

[12] “sub-sentential paraphrase” という呼び方が浸透しつつある．ここでは表層的な語の列の例を挙げているが，3.2 節で触れたように，様々な表現方法が考えうる．

木上で隣接する名詞 [38, 56, 14]，修飾語や被修飾語 [27] などが用いられてきた．

単言語パラレルコーパス (正例のみの言い換えコーパス): 同義なテキスト対が一定量あれば，統計的機械翻訳における翻訳テーブルの学習 [7, 35] と同様に，各々における部分対応を同定することにより，精度よく言い換え表現対を獲得できる．同じ文書に対する複数の人間訳 [3, 50]，数式に対する異なる説明文 [4] などが用いられてきた[13]

単言語コンパラブルコーパス: 共通する概念に対応するテキスト対から，同義の部分を同定することによって，精度よく言い換え表現対を獲得できる (e.g., (1)，(2)→(7))．同じ事柄について述べている複数の新聞社の記事 [54, 5, 60]，同じ概念や用語に対する複数の定義文 [46, 29, 64] などが用いられてきた．

異言語パラレルコーパス (対訳コーパス): 統計的機械翻訳の手法で学習した翻訳テーブルから，異なる言語において共通の訳を持つ表現を言い換えとして抽出することが考えられる [2]．統語構造を考慮することによる精緻化 [8, 65]，同期文法とみなしてのパターン化 [24]，複数の翻訳テーブルや言語資源の組み合わせによる拡張 [36] などの研究がなされてきた．このアプローチで作られた様々な言語の言い換え知識ベース [25, 43] が公開されている．

パラレルコーパスと単言語コーパスの組み合わせ: 異なる種類のコーパスを組み合わせることにより，自動獲得した言い換え知識を精緻化したり [10]，大規模化したり [22, 23] できる．

3.2 節で述べたように，頑健かつ正確な言い換え処理には膨大な言い換え知識が不可欠である．したがって，言い換え知識の自動獲得の研究においては，獲得できる (獲得した) 知識の量と質の両面から評価を行なう必要がある．河合ら [34] は，文献 [29] の手法で単言語コンパラブルコーパスから獲得した言い換え知識を，既存の語彙知識を用いた単語アラインメントに基づいて 8 種類に細分類した．そして，読みや内容語がまったく等しい「自明な」言い換え表現対のみならず，読みや内容語が異なる「非自明な」言い換え表現対を大規模かつ精度よく獲得できると述べている．Max ら [40] は，異なる方法で構築された 4 種類の単言語コンパラブルコーパスに対して人手で単語アラインメントおよび言い換え抽出を実施し，自動獲得できる言い換えの上限，および種類の分布について調査した．さらに，同じコーパスに対して 4 種類の言い換え知識獲得手法を適用し，上記の正解例を用いて性能を評価した．

[13]近年では下火になっている．4.2 節で述べるように，言い換え関係にある文の対は，異言語パラレルコーパス (対訳コーパス) のように自然に蓄積されるものではなく，大規模な言い換えコーパスの自動構築そのものが難しい問題であるためである．

4 言い換え認識技術の評価に向けて

3 節で述べたように，言い換え生成の評価は前提とする知識や適用方法に依存する上に網羅性の判定が困難であり，言い換え知識獲得についてはすでに有意義な分析がすでに行われている．他方，日本語の言い換え認識技術の分析はまだ十分に行われていない．これらをふまえ，我々は，日本語の言い換え認識タスクにおける既存の手法の到達状況や今後取り組むべき課題を明らかにするために，評価の方法論の課題を整理した．

4.1 客観的かつ精密な評価のためのシナリオ

様々な知識を組み合わせて用いるような手法 (およびシステム) の再現や処理のトレースに基づく本質的な誤り分析は，第三者には極めて困難である．また，個々の誤りが分析者によって異なる理由で説明されるようでは，解決に向けての提言は困難である．

これらをふまえ，言い換え認識技術の客観的評価を行なうために，次のシナリオを提案する．

第 1 段階. 評価に適した言い換えコーパスの構築: 複雑な事例は，解けた場合も解けなかった場合も，その理由の説明が困難である．そこで，現実世界に存在するそのような複雑な事例だけでなく，それらを可能な限りプリミティブな言い換え関係に分解した事例 [52, 33] も含む評価データを構築する．ここでは，候補とする言い換え表現対の収集・作成が課題となる (4.2 節).

第 2 段階. 必要な知識・機能の列挙: 理想的な手法が人間の処理プロセスを模倣している必要はない．ただし，誤りの傾向の分析や誤りの解消に必要な知識・機能を人間が分析・議論する上で，合意形成や説明を容易にするメタ言語は不可欠である．言い換えの種類や必要な知識・機能に関するオントロジは文献 [20] によるものが存在する．カバレージや合意形成上の有用性は不明であるが，Sammons ら [52] の成功例をふまえると，アノテーション (ここでは分類と説明) とオントロジの更新を繰り返す OntoNotes 方式 [30] で収斂できると期待できる．

第 3 段階. 既存の技術の客観的評価と課題の提言: 細かな部分問題に分解され，また各々が (ある程度) 客観的なオントロジに基づいて分類された評価データを用いると，所与の手法の性能をプロファイルできる．また，既存の語彙資源の外的

(extrinsic) な評価も可能になる．ただし，各手法に固有の問題の分析は当該手法の開発者に任せる．

4.2 言い換えコーパスが満たすべき要件

3.1 節での議論および 4.1 節のシナリオをふまえると，言い換え認識技術を的確に評価するには，次の要件を満たす言い換えコーパスを構築する必要がある．

要件 1. 分布の自然さ: 現実世界において解くべき問題の「分布」を反映することで，研究開発の方向の妥当性を保証する (cf. 言い換え (候補) 生成では多様な言い換えの網羅性)．

要件 2. 正負例のバランス: 正例と負例の境界を捉えるために両者とも必要である．特に詳細な評価を行なうためには，非自明でかつ対応する正負の境界例 (cf. Winograd Schema Challenge [37, 53]) を収録することが望ましい．

要件 3. プリミティブさ: 個々の事例をあらかじめ部分問題に分解･分類しておくことにより，手法に依拠せず客観的に達成度･誤り分析を可能にする [52, 33]．また，部分問題に対する解答に基づいて，より精密で公平な評価も可能になる．

要件 3 は，収集した事例を人間が分解することによって担保できる．要件 2 も，人間による類似例の作例である程度担保できるが，事例の候補を収集する段階で境界例が得られるのであれば，その方が望ましい．

最も難しい課題は要件 1，つまり濃度の薄さと偏りの克服である．他の多くの自然言語処理タスクは，パラレルでないコーパスを網羅的に処理するものであり，対象とするコーパスを大きくしたり，コーパスによってドメインを規定したりすることで，カバーできる現象の多様性を担保できる．異なる言語の同義表現，すなわち翻訳/対訳は，すべての言語表現に対して存在するわけではないが，局所的には日々生産されており，対訳コーパスとして明示的に蓄積することも比較的容易である．これらに対して言い換えは，場面を絞ったとしてももれなく観察できるわけではないし，任意のテキスト対はほぼ間違いなく言い換えではない．

5 トップダウンな事例候補収集の実行可能性

英語を対象とした言い換えコーパス構築の先行研究 [15, 62, 11, 63] はいずれも，トップダウンなアプローチを取っている．すなわち，何らかの仮説・手がかりに基づいて言い換えらしいテキスト対 (事例の候補) を収集し，その結果に対して人間が正負のラベルを付与することで，言い換えコーパスを構築している．このアプローチで収集できる事例の候補は，一般に図 4 のような分布になる．正負のラベル付けの人的なコストを抑えるためには，正例の割合をある程度高く[14]するための工夫が必要である．一方で，MSRP の編集距離のような行き過ぎた制約は，標本選択バイアスとなって多様性を減少させ，また，大量の自明な事例とごく少量の非自明な事例をもたらすことになる．そうならないような配慮が必要である．

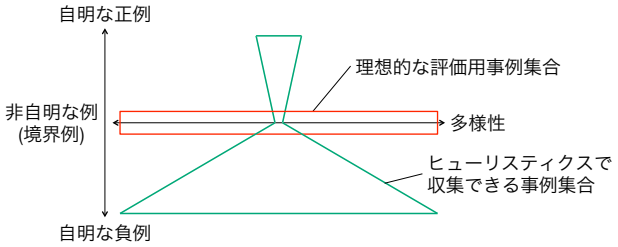

図 4: 収集できると期待される言い換え事例の範囲．

5.1 RITE-2 の評価用データの分析

我々は，上記の試みと同様にトップダウンなアプローチで構築された RITE-2 の評価データが，4.2 節の各要件を満たすか否かを調査した．まず，全 548 事例における言い換え事例 (“B”) の数は 70 件 (13%) であり，割合としては極端に少なくもないが，絶対数としては少ない．したがって，要件 1「分布の自然さ」を満たせているとは言い難い．RITE-2 のデータが元々含意関係認識の開発・評価のために構築されたことを考慮すると仕方がない．

次に，全 548 事例について，NTCIR RITE-2 タスクに参加した全 21 システムの言い換え事例 (“B”) に対する正答率を調査した[15]．3.1 節で述べた t_1 に対するトークンの重複率 (r_1) と「言い換え (“B”)」と解答したシステムの数の相関係数は 0.771 であり，r_1 に相当する情報が問題の難易度に影響していたことが窺える．言い換えの正例のみを見た場合もこの係数は 0.705 であり，次の 3 つの例 (読点の不一致は原文ママ) が示すように，共通するトークンが多いほど正解率も高かった．

(8) $r_1 = 1.00$, 正答 17, 誤答 4
t_1 太平洋戦争の敗戦に伴い、陸軍幼年学校は廃止され、解散した。
t_2 陸軍幼年学校は、太平洋戦争の敗戦に伴い廃止され、解散した。

[14] Xu ら [63] の実験では，Twitter 上で盛り上がりを見せたトピックと時間情報を用いて事例の候補を収集した場合でも正例は約 8%であった．ただし，トピックごとに出現する各語の顕現性を考慮することで，語彙の多様性は犠牲になるものの正例の割合を約 16%に，さらにトピックの選択を多腕バンディット問題と見なして最適化することで，正例の割合を約 34%まで向上させている．

[15] 参加者が提出したデータは公開されていないため，NTCIR RITE-2 のオーガナイザであった柴田，渡邉が調査を担当した．

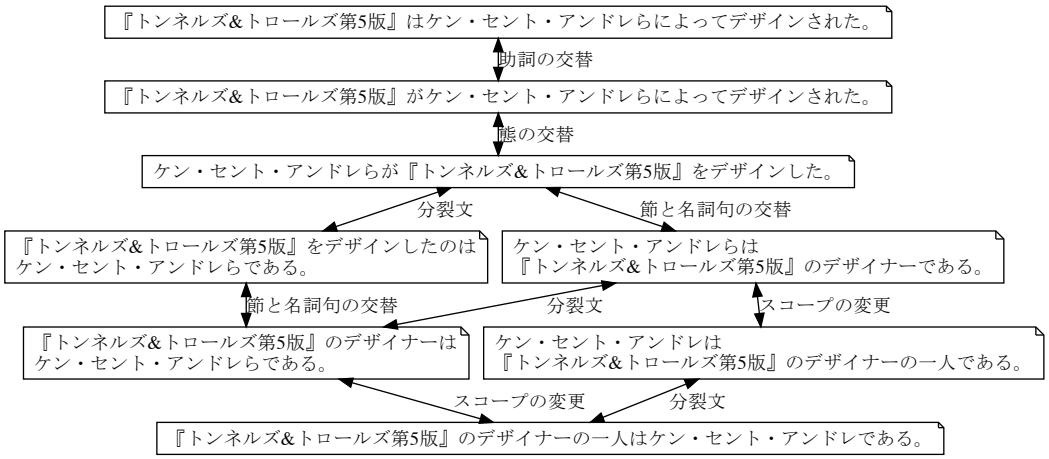

図 5: ユニットテストデータのさらなる分解の例.

(9) $r_1 = 0.75$, 正答 10, 誤答 11
- t_1 未成年者喫煙禁止法によって未成年の喫煙は禁止されている。
- t_2 未成年者喫煙禁止法は，20 歳未満の者の喫煙を禁止している。

(10) $r_1 = 0.42$, 正答 2, 誤答 19
- t_1 忍者は、漫画のキャラクターとして頻繁に登場する。
- t_2 忍者を題材とする漫画は，数多い。

また，上記の例 (8) が示すように，トークンの重複率が高いだけでなく，格要素と従属節の順序を入れ替えただけの，言い換えと呼ぶのが憚られるような自明な例も散見された．以上より，要件 2 の「正負例のバランス」は満たせていないと判断する．

要件 3 の「プリミティブさ」については，複雑な事例が多く，詳細な分析が可能でないことが文献 [33] において明らかにされている．

5.2 RITE-2 のユニットテストデータの分析

4.2 節では，収集した事例を分解することで要件 3 を満たせると述べた．このことは，Kaneko ら [33] による，RITE-2 の 2 値分類タスク[16]の評価データの分解の取り組みによって裏付けられている．この成果物は，**ユニットテストデータ**と呼ばれている．ただし，Kaneko らの作業においては分解の粒度の指標が設けられていなかったため，分解が十分でない事例も多数残っている．例えば，(Kaneko らによる) 分解済事例 (11) は，さらに図 5 のように分解できる．

(11) t_1 『トンネルズ&トロールズ第 5 版』はケン・セント･アンドレらによってデザインされた。
t_2 『トンネルズ&トロールズ第 5 版』のデザイナーの一人は、ケン・セント・アンドレである。

そこで我々は，Kaneko らが 60 事例に対して作成した 241 件の分解済事例のうち，言い換えを含むと考えられる関係をもつ 163 事例 (表 3) をさらに分解した．まず，163 事例を著者 5 人にランダムに振り分け，各自の判断で相互作用がない現象を可能な限り分解した．さらに，各自が，分解して得た個々の事例に対して，言い換えの分類体系[17]を参考にして種類名を付与した．最後に，各自の分解作業が終わった後に全員分を持ち寄り，事例のさらなる分解，種類名の統合等の整理を行った．

表 4 に示すように，163 事例のうち 60 事例が 203 事例に分解でき，それらのうち 156 件が言い換え，47 件が非言い換えであった．分解できなかった事例のうち言い換え関係にある 58 件と合わせて合計 214 件の (我々が考える限りプリミティブな) 事例における言い換えの種類の分布を表 5 に示す．一部については十分に整理・統合ができていないが，4 値分類タスクの “B” のデータよりも多くの事例を分解したため，ある程度多様な言い換えがカバーされているように見える．ただし，4.2 節の要件 1 を満たせているかどうかを判定するには至っていない．

上述の言い換えの分類体系には記載されていないが，今回分析したデータにおいては，「自明要素の明示/暗示」と名付けた現象が多数含まれていた．例えば，次の 2 つの例のように，各文を構成する要素のみではな

[16] 3.1 節で述べた 4 値のうち，“F” と “B” を含意関係あり，“C” と “I” を含意関係なしとみなすタスク．4 値分類とは異なる評価データが用いられた．

[17] http://paraphrasing.org/paraphrase.html

表 3: ユニットテストデータ (60 事例に対する 241 分解済事例) における関係の種類.

関係の種類	事例数	分析対象
synonymy:lex	10	√
hypernymy:lex	3	-
meronymy:lex	1	-
synonymy:phrase	35	√
entailment:phrase	45	√
case_alternation	7	√
modifier	42	-
nominalization	1	√
coreference	4	√
clause	14	√
relative_clause	8	√
transparent_head	1	√
list	3	-
scrambling	15	√
inference	2	√
implicit_relation	18	√
apposition	1	√
temporal	1	√
spatial	1	√
disagree:lex	2	-
disagree:phrase	25	-
disagree:modality	1	-
disagree:temporal	1	-
Total	241	

表 4: ユニットテストデータの再分解の結果.

再分解	分解済事例数	言い換え	非言い換え	合計
なし	103	58	45	103
あり	60	156	47	203
合計	163	214	92	306

く,「映画」という語の特質構造の知識,「化身ラマ制度」が「法主の選任」のための制度であるという世界知識がなくては言い換えであることを判定できないような事例も含まれていた.

(12) a. 『ステンカ・ラージン』はウラジミール・ロマシコフが監督、ワシーリ・ゴンチャロフが脚本**の**映画だ。
b. 『ステンカ・ラージン』はウラジミール・ロマシコフが監督、ワシーリ・ゴンチャロフが脚本**で制作された**映画だ。

(13) a. カルマ・カギュ派が、化身ラマ制度を初めて**法主の選任に**採用した。
b. カルマ・カギュ派が、化身ラマ制度を初めて採用した。

トップダウンに事例の候補を収集するアプローチでは，このような現象を含む本質的に難しい事例も得られるが，自明な例と同じ「1 事例」とみなしていては解決の糸口が見えない．また，評価も公正には行えない．これに対して，各事例をプリミティブな事例に分解することにより，上記のような分析，また部分点を考慮した評価が可能になる．

表 5: ユニットテストデータの再分解によって得られた言い換えの種類の分布.

言い換えの種類	分解済	新規獲得	合計
名詞/名詞	1	7	8
名詞句/名詞句	0	2	2
動詞/動詞	0	6	6
動詞/動詞句	1	2	3
動詞句/動詞句	1	2	3
副詞/副詞	1	1	2
略記	0	1	1
表記の揺れ	0	2	2
助詞の交替	2	31	33
助動詞	0	4	4
テンス・アスペクト表現の正規化	0	2	2
機能表現	0	3	3
複合名詞化	1	1	2
同格表現の異形	1	0	1
括弧・同格	1	0	1
括弧の付加/削除	0	5	5
並列名詞句の入れ替え	2	1	3
並列動詞句の入れ替え	0	2	2
格要素の語順の変更	4	9	13
数量詞の移動	0	1	1
主題の交替	9	10	19
態の交替	5	6	11
相互格の交替	2	0	2
機能動詞構文	1	4	5
動詞句/名詞句	0	1	1
文法カテゴリを変える言い換え	0	3	3
所有–存在	0	1	1
地名–存在	1	1	2
分裂文	0	2	2
節/名詞句	0	3	3
節の統合/分割	1	0	1
節の連体修飾節化	2	1	3
節をまたぐ言い換え	0	2	2
文の統合/分割	0	4	4
共参照表現による置換	3	5	8
コピュラ文の主辞の削除/挿入	1	3	4
自明要素の明示/暗示	15	9	24
説明の省略	2	3	5
数量詞の省略	0	1	1
非制限的説明の除去	0	2	2
スコープの変更	0	1	1
句読点	0	4	4
未分類	1	8	9
合計	58	156	214

今回の事例の分解作業を通じて負例も得られたが，言い換えあるいは一方向の含意の事例のみを分解の対象としたため，相対的に少量であった．また，このようにして得られた負例が，正例との境界に存在する紛らわしい事例であるとは限らない．やはり人間による境界例の作例は不可欠であろう．

6 おわりに

日本語を対象とした言い換え認識手法の研究はまだ包括的には行われておらず，既存の手法の到達状況や今後取り組むべき課題を明らかにするためには，まずは評価の方法論について検討する必要がある．本稿では，英語における同タスクの研究動向および NTCIR RITE-2 において構築された日本語におけるテキスト間含意関

係認識技術の評価データの分析をふまえ，言い換え認識タスクの評価のシナリオを提案した．そして，最初に取り組むべき課題である，評価に適した言い換えコーパスの構築に向けて，コーパスが満たすべき要件を,「分布の自然さ」,「正負例のバランス」,「プリミティブさ」の 3 つに整理した．さらに，RITE-2 で構築されたユニットテストデータを用いて，所与の言い換え事例を相互作用のないプリミティブな事例 (部分問題) に分解できること，それによって，システムの出力に対する主観的な分析ではなく，事例に由来する (システム横断的という意味で) 客観的な評価が可能になる見通しを示した．

今後まず取り組むべき課題は，上記の要件を満たす評価用コーパスの構築である．1 つ目の「自然な分布」については，現時点では明確な解決法は存在しないが，他の言語を題材として提案された近年の手法を含む複数の言い換えコーパス構築手法を比較しながら検討を進めることが重要である．これまでの我々の活動の延長で，RITE-2 の分析結果を公開することは可能であるが，その場合も，コミュニティをミスリードしないよう,「自然な分布」について十分に検討することが先決であると認識している．

謝辞: 本研究の一部は科研費若手研究 (B)（課題番号：25730139，代表：藤田篤）の支援を受けた．情報通信研究機構の飯田龍氏から，本ワークショップにおける議論の題材として 2 節を追記することが有意義であるとの助言を受けた．ここに記して感謝する．

参 考 文 献

[1] E. Agirre, D. Cer, M. Diab, and A. Gonzalez-Agirre. SemEvel-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics* (∗*SEM*), pp. 385–393, 2012.

[2] C. Bannard and C. Callison-Burch. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 597–604, 2005.

[3] R. Barzilay and K. R. McKeown. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 50–57, 2001.

[4] R. Barzilay and L. Lee. Bootstrapping lexical choice via multiple-sequence alignment. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp. 164–171, 2002.

[5] R. Barzilay and L. Lee. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of the 2003 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL*), pp. 16–23, 2003.

[6] R. Bhagat and D. Ravichandran. Large scale acquisition of paraphrases for learning surface patterns. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 161–170, 2008.

[7] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, Vol. 19, No. 2, pp. 263–311, 1993.

[8] C. Callison-Burch. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp. 196–205, 2008.

[9] C. Callison-Burch, I. Dagan, C. Manning, M. Pennacchiotti, and F. M. Zanzotto, editors. *The 2009 Workshop on Applied Textual Inference* (*TextInfer*), 2009. ACL/IJCNLP 2009 Workshop.

[10] T. P. Chan, C. Callison-Burch, and B. V. Durme. Reranking bilingually extracted paraphrases using monolingual distributional similarity. In *Proceedings of the Workshop on Geometrical Models of Natual Language Semantics* (*GEMS*), pp. 33–42, 2011.

[11] D. L. Chen and W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 190–200, 2011.

[12] T. Cohn and M. Lapata. Sentence compression beyond word deletion. In *Proceedings of the 22nd International Conference on Computational Linguistics* (*COLING*), pp. 137–144, 2008.

[13] I. Dagan, O. Glickman, and B. Magnini. The PASCAL recognizing textual entailment challenge. In *Recognizing Textual Entailment Challenge in PASCAL - Pattern Analysis, Statistical Modelling and Computational Learning*, pp. 1–8, 2005.

[14] S. De Saeger, K. Torisawa, M. Tsuchida, J. Kazama, C. Hashimoto, I. Yamada, J. Oh, I. Varga, and Y. Yan. Relation acquisition using word classes and partial patterns. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp. 825–835, 2011.

[15] B. Dolan, C. Quirk, and C. Brockett. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th International Conference on Computational Linguistics* (*COLING*), pp. 350–356, 2004.

[16] M. Dras. A meta-level grammar: Redefining synchronous TAG for translation and paraphrase. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 80–88, 1999.

[17] A. Finch, Y.-S. Hwang, and E. Sumita. Using machine translation eveluation techniques to determine sentence-level semantic equivalence. In *Proceedings of the 3rd International Workshop on Paraphrasing* (*IWP*), pp. 17–24, 2015.

[18] 藤田篤, 乾健太郎, 乾裕子. 名詞言い換えコーパスの作成環境. 電子情報通信学会技術研究報告, TL2000-32, pp. 53–60, 2000.

[19] 藤田篤, 乾健太郎. 語彙・構文的言い換えにおける変換誤りの分析. 情報処理学会論文誌, Vol. 44, No. 11, pp. 2826–2838, 2003.

[20] 藤田篤, 乾健太郎, 松本裕治. 言い換え知識の類型化と例文集構築の試み. 言語処理学会第 10 回年次大会発表論文集, pp. 420–423, 2004.

[21] A. Fujita and K. Inui. A class-oriented approach to building a paraphrase corpus. In *Proceedings of the 3rd International Workshop on Paraphrasing* (*IWP*), pp. 25–32, 2005.

[22] A. Fujita, P. Isabelle, and R. Kuhn. Enlarging paraphrase collections through generalization and instantiation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (*EMNLP-CoNLL*), pp. 631–642, 2012.

[23] A. Fujita and P. Isabelle. Expanding paraphrase lexicons by exploiting lexical variants. In *Proceedings of Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (*NAACL-HLT*), 2015. (to appear).

[24] J. Ganitkevitch, C. Callison-Burch, C. Napoles, and B. V. Durme. Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp. 1168–1179, 2011.

[25] J. Ganitkevitch and C. Callison-Burch. The multilingual paraphrase database. In *Proceedings of the 9th International Conference on Language Resources and Evaluation* (*LREC*), pp. 4276–4282, 2014.

[26] 後藤功雄, 熊野正, 田中英輝. 一般のニュースからやさしい日本語ニュースへの書き換えの分析. 言語処理学会第 20 回年次大会発表論文集, pp. 15–18, 2014.

[27] M. Hagiwara, Y. Ogawa, and K. Toyama. Selection of effective contextual information for automatic synonym acquisition. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics and the 21st International Conference on Computational Linguistics* (*COLING-ACL*), pp. 353–360, 2006.

[28] Z. Harris. Distributional structure. *Word*, Vol. 10, No. 23, pp. 146–162, 1954.

[29] C. Hashimoto, K. Torisawa, S. De Saeger, J. Kazama, and S. Kurohashi. Extracting paraphrases from definition sentences on the Web. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 1087–1097, 2011.

[30] E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL*) *Short Papers*, pp. 57–60, 2006.

[31] 乾健太郎, 藤田篤. 言い換え技術に関する研究動向. 自然言語処理, Vol. 11, No. 5, pp. 151–198, 2004.

[32] 梶原智之, 山本和英. 日本語の語彙平易化評価セットの構築. 言語処理学会第 21 回年次大会発表論文集, 2015.

[33] K. Kaneko, Y. Miyao, and D. Bekki. Building Japanese textual entailment specialized data sets for inference of basic sentence relations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 273–277, 2013.

[34] 河合剛巨, 橋本力, 鳥澤健太郎, 川田拓也, 佐野大樹. 定義文から自動獲得した言い換えフレーズペアの分析. 言語処理学会第 18 回年次大会発表論文集, pp. 421–424, 2012.

[35] P. Koehn. *Statistical Machine Translation*. Cambridge University Press, 2009.

[36] S. Kok and C. Brockett. Hitting the right paraphrases in good time. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (*NAACL-HLT*), pp. 145–153, 2010.

[37] H. J. Levesque. The Winograd Schema Challenge. In *Proceedings of the 10th International Symposium on Logical Formalization on Commonsense Reasoning, AAAI 2011 Spring Symposium*, 2011.

[38] D. Lin and P. Pantel. Discovery of inference rules for question answering. *Natural Language Engineering*, Vol. 7, No. 4, pp. 343–360, 2001.

[39] Y. Marton. Distributional phrasal paraphrase generation for statistical machine translation. *ACM Transactions on Intelligent Systems and Technology*, Vol. 4, No. 3, 2013.

[40] A. Max and A. V. Houda Bouamor. Generalizing subsentential paraphrase acquisition across original signal type of text pairs. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (*EMNLP-CoNLL*), 2012.

[41] D. McCarthy and R. Navigli. The english lexical substitution task. *Language Resource and Evaluation*, Vol. 43, No. 2, pp. 139–159, 2009.

[42] I. Mel'čuk and A. Polguère. A formal lexicon in Meaning-Text Theory (or how to do lexica with words). *Computational Linguistics*, Vol. 13, No. 3-4, pp. 261–275, 1987.

[43] M. Mizukami, G. Neubig, S. Sakti, T. Toda, and S. Nakamura. Building a free, general-domain paraphrase database for Japanese. In *Proceedings of the 17th Oriental COCOSDA Conference*, 2014.

[44] 水野淳太, E. Nichols, 渡邉陽太郎, 村上浩司, 松吉俊, 大木環美, 乾健太郎, 松本裕治. 言論マップ生成技術の現状と課題. 言語処理学会第 17 回年次大会発表論文集, pp. 49–52, 2011.

[45] 村松祐希, 山本和英. 語彙知識を用いた日本語テキスト含意認識評価セット構築と認識実験. 言語処理学会第 16 回年次大会発表論文集, pp. 514–517, 2010.

[46] 村田真樹, 井佐原均. 複数の辞書の定義文の照合に基づく同義表現の自動獲得. 自然言語処理, Vol. 11, No. 5, pp. 135–149, 2004.

[47] 名取芙美香, 松吉俊, 福本文代. 含意認識タスクに関するかき混ぜ文対データの構築. 言語処理学会第 20 回年次大会発表論文集, pp. 745–748, 2014.

[48] 小谷通隆, 柴田知秀, 中田貴之, 黒橋禎夫. 日本語 Textual Entailment のデータ構築と自動獲得した類義表現に基づく推論関係の認識. 言語処理学会第 14 回年次大会発表論文集, pp. 1140–1143, 2008.

[49] M. Paşca and P. Dienes. Aligning needles in a haystack: Paraphrase acquisition across the Web. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing* (*IJCNLP*), pp. 119–130, 2005.

[50] B. Pang, K. Knight, and D. Marcu. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of the 2003 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL*), pp. 102–109, 2003.

[51] C. Quirk, C. Brockett, and W. Dolan. Monolingual machine translation for paraphrase generation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp. 142–149, 2004.

[52] M. Sammons, V. G. V. Vydiswaran, and D. Roth. "ask not

what textual entailment can do for you...". In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 1199–1208, 2010.

[53] 柴田知秀, 小浜翔太郎, 黒橋禎夫. 日本語 Winograd Schema Challenge の構築と分析. 言語処理学会第 21 回年次大会発表論文集, 2015.

[54] Y. Shinyama, S. Sekine, K. Sudo, and R. Grishman. Automatic paraphrase acquisition from news articles. In *Proceedings of the 2002 Human Language Technology Conference* (*HLT*), 2002.

[55] R. Socher, E. H. Huang, J. Pennington, A. Y. Ng, , and C. D. Manning. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems* (*NIPS*), 2011.

[56] I. Szpektor, H. Tanev, I. Dagan, and B. Coppola. Scaling Web-based acquisition of entailment relations. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp. 41–48, 2004.

[57] T. Takahashi, T. Iwakura, R. Iida, A. Fujita, and K. Inui. KURA: A transfer-based lexico-structural paraphrasing engine. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium* (*NLPRS*) *Workshop on Automatic Paraphrasing: Theories and Applications*, pp. 37–46, 2001.

[58] 宇高邦弘, 山本和英. 複数の客観的手法を用いたテキスト含意関係評価セットの構築. 言語処理学会第 17 回年次大会発表論文集, pp. 627–630, 2011.

[59] M. Vila, M. M. Antònia, and H. Rodríguez. Is this a paraphrase? what kind? paraphrase boundaries and typology. *Open Journal of Modern Linguistics*, Vol. 4, pp. 205–218, 2014.

[60] R. Wang and C. Callison-Burch. Paraphrase fragment extraction from monolingual comparable corpora. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora*, pp. 52–60, 2011.

[61] Y. Watanabe, Y. Miyao, J. Mizuno, T. Shibata, H. Kanayama, C.-W. Lee, C.-J. Lin, S. Shi, T. Mitamura, N. Kando, H. Shima, and K. Takeda. Overview of the recognizing inference in text (RITE-2) at NTCIR-10. In *Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies*, pp. 385–404, 2013.

[62] S. Wubben, A. van den Bosch, E. Krahmer, and E. Marsi. Clustering and matching headlines for automatic paraphrase acquisition. In *Proceedings of the 12th European Workshop on Natural Language Generation* (*EWNLG*), pp. 122–125, 2009.

[63] W. Xu, A. Ritter, C. Callison-Burch, W. B. Dolan, and Y. Ji. Extracting lexically divergent paraphrases from Twitter. *Transaction of the Association for Computational Linguistics* (*TACL*), Vol. 2, pp. 435–448, 2014.

[64] Y. Yan, C. Hashimoto, K. Torisawa, T. Kawai, J. Kazama, and S. De Saeger. Minimally supervised method for multilingual paraphrase extraction from definition sentences on the web. In *Proceedings of Human Language Technologies: The 2013 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (*NAACL-HLT*), pp. 63–73, 2013.

[65] S. Zhao, H. Wang, T. Liu, and S. Li. Extracting paraphrase patterns from bilingual parallel corpora. *Natural Language Engineering*, Vol. 15, No. 4, pp. 503–526, 2009.

[66] Z. Zhu, D. Bernhard, and I. Gurevych. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics* (*COLING*), pp. 1353–1361, 2010.