言い換えエンジンKURAを用いた節内構造および機能語相当表現レベルの言い換え

3H−3

飯田龍† 徳永泰浩† 乾健太郎† 衛藤純司‡

†九州工業大学情報工学部知能情報工学科 ‡インターナショナル・ランゲージウェア

1 はじめに

テキストの言い換えは，様々な自然言語処理の応用分野に貢献できる重要な技術である．近年，これを応用からある程度独立した要素技術として捉え，問題の性質や実現方法を追究する試みも見られ，研究者の関心が高まってきた[2]．言い換えは，翻訳の一種と見なせるが，異言語間翻訳とは異なる側面もあり，その性質の解明はまさにこれからの課題である．

以上の背景から，我々は，構文構造変換をベースとする言い換えの効率的な実験環境KURAを開発した[4, 6]．KURAでは，言い換えに必要な知識を，言い換えを駆動する構造変換モデルと構造変換結果を修正・棄却するための言語モデルに分け，それぞれの知識を統一的かつ宣言的な依存構造書き換え規則として記述する．修正・棄却機構を導入し，知識をモジュール化することにより，言い換え知識全体の複雑さ（開発コスト）を抑えるのがねらいである．文献[4]で報告したように，KURAの初期的な設計・実装がある程度収束を見た現在，次に取り組むべき課題は，(a) 実際にKURAを用いて，多様な言い換えに必要な知識を宣言的な記述形式で蓄積すること，また (b) その過程で得られる知見をKURAのアーキテクチャにフィードバックすることである．

一口に言い換えと言っても様々な種類があり，我々が当面の対象としている語彙・構造的 (lexico-structural) な言い換えに限っても，節間（文全体）の構造を組み換える言い換えから，節内に閉じた言い換え，局所的な語の置き換えまで様々である[3]．本稿では，これらのうち節内構造レベルおよび機能語相当表現レベルの言い換えを取り上げる．この種の言い換えは，(a) 変換のパタンがクローズドなクラスに抑えられ，ある程度人手で書き尽せる，(b) 主辞交替や格交替，副詞・副助詞の呼応など，様々な興味深い言語現象を伴う場合が少くない，(c) 節間構造の組み換えに比べると文脈の影響が比較的少い，(d) 言い換えの直接的な応用の一つである文の簡単化（平易化）[3]に効果的である，といった性質を持っており，KURAの使用実験の例題として適当であると判断した．

以下，節内構造レベル，機能語相当表現レベルのそれぞれの言い換えについて，これをKURA上に実装した結果について報告し，得られた知見のいくつかを論じる．

2 節内構造レベルの言い換え

2.1 実験

節内構造レベルの言い換えの例題として，次の例ような，否定と副助詞が呼応する表現の言い換えを取り上げた．これらの言い換えは，主辞交替を伴う言い換えの代表的なものであり，副詞と否定の呼応，副助詞と格助詞の相互作用，動詞の格交替といった様々な言語現象を引き起こす点で興味深い例題である．

(1) りんごしか食べない→りんごだけ食べる（副助詞が否定「ない」を含む述語に係る）

(2) お酒を飲まない人は彼だけだ→彼以外は全員がお酒を飲む（否定「ない」を含む関係節の被修飾名詞が副助詞を伴う述語に係る）

(3) 彼ほどお酒を飲む人はない→彼が一番お酒を飲む（副助詞を含む関係節の被修飾名詞が否定「ない」を含む述語に係る）

まず，京大コーパス[5]約4万文から，「副助詞が否定『ない』を含む述語に係る」など，上の3つの条件のいずれかを満たす文を機械的に抽出し，約2000文を得た．次に，得られた2000文から副助詞が「ない」と呼応している文，230文を人手で抽出し，これを人手で言い換えて422件の言い換え事例を作成した．ここで，「副助詞が『ない』と呼応する」とは，その文から「ない」を取り除くと，副助詞が存続できないか，あるいは意味が変ってしまう場合を指す．次に，作成した言い換え事例に基づいて，それらを実現する構造変換規則177規則を作成し，KURA上に実装した．以下は，実装した変換規則の例である*1．

(4) `N-だけV-する -> N-以外はV-する．`

(5) `N-しかV-しない -> V-するのはN-だけ．`

言い換え生成実験では，まず上の177規則を前述の2000文に適用し，377件の言い換えを生成した．ただし，上の規則の例からも容易に想像できるように，実装した変換規則は適用条件や変換処理の記述が不完全 (underspecified) であり，これらを単純に適用するだけでは大量の誤りが生じる．そこで，言い換え結果の誤りを人手で分析し，これを修正・棄却するための規則（言語モデル）を人手で作成した．以下は，作成した修正・棄却規則の例である．

(6) `ほどだけV-する -> ほどV-するだけ．`

(7) `N1-だけN2 -> N1-だけというN2．`

訓練事例に対する現時点のパフォーマンスは，適合率で62%（160件/258件）である．ただし，上述の377件のうち，言い換えエンジンのバグ，あるいは言い換えによって従属節との修辞的関係が壊れる場合など，今回の実験のスコープを越える誤りを除いた258件を評価の対象とした．

2.2 考察

本実験が手作業による規則作成に基づくものであることを考えると，適合率62%は訓練事例に対するパフォーマンスとしては決して高くない．これは，主として，「〜だけでない」のような頻出表現が〈存在〉（「〜がある」）の否定の意味と〈コピュラ〉（「〜は〜だ」）の否定の意味の間で多義であり，この多義性をうまく扱えていなかったことが原因である．この問題による誤りは誤り全体の67%あった．

残りの誤りは，「一つ」や「今」といった数詞，副詞的名詞に対する格助詞補完の誤り，テンス・アスペクト表現の修

Exploration of clause-structural and funtion-expressional paraphrasing using KURA

Iida Ryu†, Tokunaga Yasuhiro†, Inui Kentaro† and Etoh Junji‡
†Kyushu Institute of Technology ‡International Langageware
{r_iida,y_toku,inui}@pluto.ai.kyutech.ac.jp

*1 N, Vは，それぞれ名詞，動詞を表すマクロ記号である．構造変換規則および修正・棄却規則の仕様については，文献[4, 6]を参考にされたい．

正誤りなどが目立った．これらの問題は統語・意味的な比較的一般性の高い制約によって扱える可能性があり，言語モデルを洗練・拡張するための良い例題になると考えられる．

また，これらの作業を通じて，KURAの知識表現能力の問題もいくつか明らかになった．たとえば，次の規則は(9)のような適格な言い換えを生成する反面，現状では(10)のような意味的に不適格な言い換えも生成してしまう．

(8) `N-だけ V-ない -> N-以外は V-する．`

(9) 私だけ仕事をしなかった → 私以外は仕事をした

(10) 私だけさぼっていて，仕事をしなかった → * 私以外はさぼっていて仕事をした

この問題は，規則を適用する際に「さぼっていて，仕事をしなかった」という並列構造を考慮していないことが原因である．逆に，そういった制御を記述する枠組みが現在のKURAにないことが問題であるとも言える．

また，現在の仕様では，原文のうち，構造変換によって破壊された部分の情報に修正・棄却規則から直接アクセスすることができない．今回の実験では，言い換えの前後で意味が保存されているかを修正・棄却時にチェックしたい場合に，この制約が問題になる場合がいくつか見られた．これについては，今後何らかの工夫が必要である．

3 機能語相当表現レベルの言い換え

3.1 実験

機能語相当表現とは，いわゆる「表現文型」という術語で一括りにされる助詞・助動詞相当表現である．このレベルでの言い換えは節内・節間レベルの言い換えに比べ語彙的な性格が強い．これらの表現の中には「涙ながらに→涙を流しながら」，「〜しがてら→〜するのを兼ねて」，「〜のきらいがある→〜の傾向がある」のように局所的な情報を参照するだけで言い換えられるものも多い．

本研究では，日本語能力試験出題基準「1・2級の機能語の類のリスト」[7]に準拠した3種類の日本語教材[8, 9, 10]に掲載されている解説や例文をもとに，279個の言い換え規則をKURAに実装した．また同教材から例文185文を収集した．さらに，言い換え規則と例文に対する言い換え事例754件について誤り分析を行い，言語モデルとして53規則を作成した．以下は，その例である．

(11) `N(sem_class!:2351_色) 一色 -> reject.`

(12) `だけならず -> reject.`

(11)の規則は〈名詞〉でかつ意味素性が〈色〉ではない形態素が「一色」に係る句が存在した場合，言い換え結果を棄却するということを意味する．(12)は言い換えられた文に「だけならず」という句が存在した場合，言い換え結果を棄却することを意味する．

以上のような言語モデルを実装した結果，上の254件の訓練事例に対しては，再現率，適合率ともに92% の精度を得るに至った．最後に，京大コーパス[5] 4万文の一部6678文に対してオープンテストを行った．この時の適用箇所は235箇所で，適合率は39 %とかなり低い値となった．

3.2 考察

オープンテストにおける誤りの原因として，まず機能語相当表現の多義性があげられる．訓練事例では観察できなかった多義性が少なくなかったかったため，規則がそれらに対応できていなかった．これについては，機能語相当表現がどのような多義性を持つかをさらに調査し，それを解消する意味解析を取り入れる必要がある．

また，今回の実験では，規則の訓練時に，個々の誤り事例に対して個別的な対処しかしなかった．したがって，得られた言語モデルの規則は，規則(12)の例のように，個々の言い換えにかなり特化した一般性の低いものが大半である．このため，オープンなデータに対してはほとんど修正・棄却の機能が働かなかった．これはある程度最初から予測された事態である．今後は，もうしばらく個別的な修正・棄却規則の収集を進め，それらが十分に蓄積された時点で分類・整理し，既存の辞書やコーパスを利用して規模の拡大をはかる．

今回の実験では，事例から言い換え後の誤りパタンを検出し，それらを修正・棄却するように言語モデルを作成した．しかし，誤りパタンは自然に存在する言語データからは獲得できず，また適格なパタンよりもさらに多様性が大きい恐れもあるので，誤りパタンの記述に頼る方法には拡張性に限界がある．(11)のように正しいパタンを規則として蓄積する方がコスト的にも，知識の網羅性を考えた上でも優れていると考えられるが，具体的にどのようにそれを実現すればよいかを検討するにはさらに事例の蓄積が必要である．

4 おわりに

本稿では，節内構造レベルおよび機能語相当表現レベルの言い換えに必要な知識を言い換え実験環境KURA上でハンドコーディングする試みについて報告した．

言い換えでは，解析技術の高度化に加えて，変換規則の不備を吸収するロバストな言語モデルの開発が鍵であると我々は考えている．そのためには，(a) 誤り分析を通じて言語モデルに必要な知識の種類をボトムアップに収集し，(b) 個々の種類についてその規模をコーパスや辞書からの知識獲得によって拡大する，というプロセスを繰り返す必要がある．本稿で報告した実験は，(a) の作業に対する試みの一つである．KURAではこの (a) の作業を効率的に支援することを目的として統一的な知識記述環境を提供しているが，今回の実験を通して問題点もいくつか明らかになってきた．

今後は，当面今回取り上げた言い換えの例題による実験を継続し，上の (a)，(b) の作業を引き続き繰り返すとともに，実験環境KURAの洗練を進める予定である．また，構造変換パタンの自動獲得の研究（e.g. [1]）にも取り組む．

参考文献

[1] Barzilay, R. and McKeown, K. Extracting paraphrases from a parallel corpus. *ACL-EACL'2001*, pp. 50–57, 2001.

[2] 言語処理学会．言語処理学会第7回年次大会ワークショップ論文集，2001.

[3] 乾健太郎．コミュニケーション支援のための言い換え．言語処理学会第7回年次大会併設ワークショップ，2001.

[4] 岩倉友哉，高橋哲朗，飯田龍，乾健太郎．KURA：統一的かつ宣言的知識記述に基づく言い換えエンジン．情報処理学会全国大会，デ–19，2001.

[5] 黒橋禎夫，長尾 眞．京都大学テキストコーパス・プロジェクト．言語処理学会第3回年次大会発表論文集，pp. 115–118，1997.

[6] 高橋哲朗，岩倉友哉，飯田龍，乾健太郎．KURA：統一的かつ宣言的記述法に基づく言い換え知識の開発環境．電子情報通信学会思考と言語研究会，2001.

[7] 国際交流基金．日本語能力試験 出題基準．凡人社，1994.

[8] 白寄まゆみ他．日本語能力試験対応 文法問題集1級・2級．桐原ユニ，1998.

[9] 友松悦子他．どんな時どう使う 日本語表現文型500．アルク，1996.

[10] 植木香他 / アジア学生文化協会留学生日本語コース．完全マスター1級/2級 日本語能力試験文法問題対策．スリーエーネットワーク，1997.