3.5 言い換え技術

―同じ意味を持つ異なる言語表現を扱う―

藤田 篤（情報通信研究機構）

言い換えの認識と生成

たとえば，「重傷を負う恐れがある」と「大ケガをしてしまうかもしれない」のように，自然言語には，同じ意味内容を表す（同じ言語の）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』と言う．

Webの検索や質問応答などでは，検索対象の文書とまったく同じ言い回しでユーザが検索したり質問したりするとは限らない．柔軟な照合のためには，任意の2つの言語表現が同じ意味を持つかどうかを判定する処理（**図-1**，言い換え認識）が必要である．一方，子どもや語学学習者とのコミュニケーションにおいては，彼らに合わせて表現を簡単にする必要があるし，音声対話システムの発話生成時には複雑な構文や同音異義語をなるべく避けたい．このように，与えられた言語表現を目的に応じて変換する処理（**図-2**，言い換え生成）に対するニーズも多い．

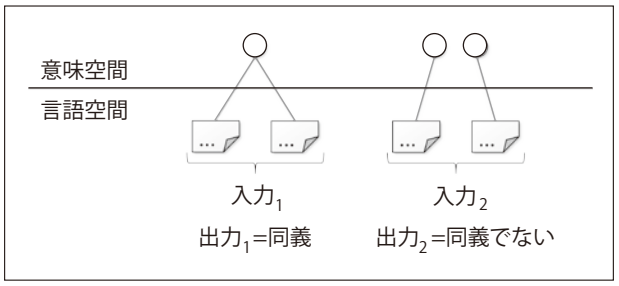

図-1　言い換え認識

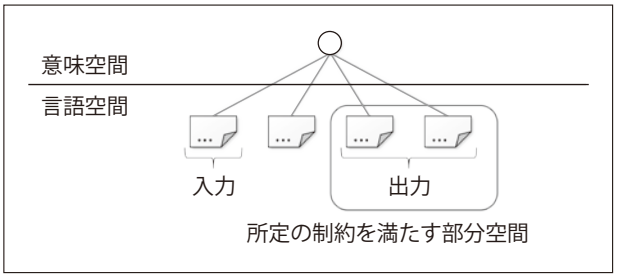

図-2　言い換え生成

一般的な解法

言い換え認識も言い換え生成も，実際には，意味空間を陽には経由せず，入力された言語表現に対する基礎的な言語解析と辞書や大規模なテキストデータなどの言語資源を用いて実現されてきた．

言い換え認識技術は，正解データ（2つの言語表現とそれらに対して人間が同義か否かを判定したデータ）を用意することで，性能を自動評価しながらの開発が可能になる．たとえば英語については，Microsoft Research Paraphrase Corpus（MSRP）の登場により研究が発展した☆1．既存の手法はすべて，(1) 入力された2つの言語表現の間の部分的な対応関係（たとえば「重傷」と「大ケガ」が同義であること）や差分を同定し，(2) そのような各特徴の重要度を正解データに基づいて推定しておき，(3) 新しい入力の同義性はその重要度に基づいて判定する，というものである．

既存の言い換え生成手法は，(1) 入力された言語表現の一部を同義語辞書などを用いて書き換えて複数の候補を生成し，(2) それらの各々を目的や制約に照らしてスコアリングして最適なものを選ぶ，という枠組みでまとめられる．ただし，言い換え生成の場合は一般に妥当な解は複数存在し，事前にすべてを列挙することも困難であるため，出力を得るたびに人間による評価が必要である．

言い換え認識と言い換え生成の両方で，「重傷」と「大ケガ」のような同義表現に関する膨大な知識が不可欠である．近年，そのような知識をテキストデータから自動獲得する研究が盛んに行われている．

☆1 http://aclweb.org/aclwiki/index.php?title=Paraphrase_Identification_(State_of_the_art)

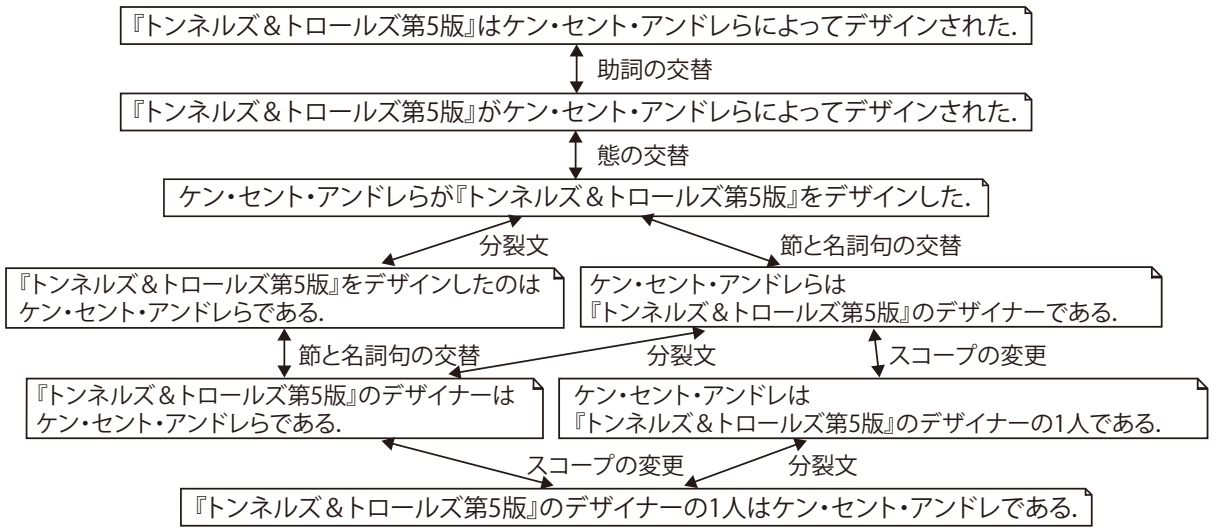

図 -3　言い換え関係にある文のネットワーク表現（最小限の言い換え関係のみを辺としている）

単語のみでなく，イディオムや一般のフレーズも対象として，数千万～数億件の言い換え表現対が蓄積されつつある☆2.

研究コミュニティとしての課題

言い換え技術をさらに発展させるために，少なくとも次の 2 つの課題に，研究コミュニティとして取り組む必要がある.

第 1 に，言い換えという現象を丁寧に分析・整理する必要がある．言い換えとみなせる多様な現象の全体像は明らかになっていないし，言い換えとみなすべき現象とそうでない現象との境界は目的によって異なる．目的に応じて重要視すべき現象を見極め，それに適した手法を検討するためにも，現象の体系化が必要である.

第 2 に，評価方法の確立を挙げる．言い換え認識においては，英語における MSRP が多くの研究者を呼び込んだが，一方で，カバーする現象の多様性や，正負のデータのバランスに問題があると指摘されている．日本語を対象とする場合も含めて，今後開発・評価データを作成する際は，これらの点に注意すべきである．また，公平な評価を可能にしたり，問題の所在を分析したりするには，**図 -3** のように，最小限の言い換えのみからなる表現の対を蓄積することも必要だろう．言い換え生成においては，効率的な評価方法を確立する必要がある．同様にテキストを生成する機械翻訳や要約における試みが参考になる.

End-to-end のモデル化

ほかの多くの自然言語処理技術と同様に，言い換え認識にも深層学習が用いられ，言語表現の再帰性や意味の構成性のモデル化が試みられるようになってきた．言い換え生成についても，同様のモデルを用いることで，語句を適切な順序で次々と生成できる可能性がある．膨大な同義表現の知識ベースを陽に持ち，証拠を丁寧に組み合わせる従来のアプローチに取って代わるのか？　両者は融合できるのか？今後の展開から目が離せない.

(2015 年 9 月 25 日受付)

藤田 篤（正会員） atsushi.fujita@nict.go.jp
2005 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了．博士（工学）．京都大学産学官連携研究員，名古屋大学助教，公立はこだて未来大学准教授を経て，2014 年より情報通信研究機構主任研究員．自然言語処理の研究に従事.

☆2　http://paraphrase.org/