SANDGLASS: 両言語換言機構を基軸とする音声翻訳

山本 和英　白井 諭　坂本 仁　張 玉潔
E-mail: {yamamoto, shirai, msakamo, yzhang}@slt.atr.co.jp
ATR 音声言語通信研究所

概要

単言語処理に処理の重点を移し、原言語と目的言語の双方で換言処理 (paraphrasing) を行なうことで変換処理の負荷を最小限に抑えた音声翻訳システム SANDGLASS の基本設計を述べ、翻訳方式の議論を行なう。

1　動機と提案の概略

英語が翻訳対象言語でない場合、例えば中国語と日本語の間での音声翻訳について述べる。これらの言語間においては日英や英日とは異なり対訳コーパス (bilingual corpus) の入手は容易ではないため、対訳コーパスを利用した翻訳手法は有力な手法とは言えない。特に音声翻訳の場合はこの傾向が顕著であり、将来に渡って複数言語で対応づけされた大量の音声言語資源の入手は現実的ではないため、コーパス利用の手法は実用的ではない。一方、言語によっては二言語話者を確保することさえ容易ではないため、手作業による翻訳知識の構築も比較的困難である。日英／英日翻訳とは異なるこのような状況下で、我々はどのような音声翻訳モデルを考えるのが現実的かについて議論する。

幸い、日本語もしくは中国語に限定すれば、近年単言語コーパスの入手は容易となり、単言語話者に作業依頼することも比較的容易である。もし、これを有効活用して原言語および目的言語で単言語処理を十分に行なうことができれば、二言語に直接関わる知識及び処理を低減した翻訳機構を構築できるのではないかと考えた。すなわち、変換処理に基軸を置いた機械翻訳モデルを転換し、原言語と目的言語の両者の換言処理を翻訳機構の基軸に置いた機械翻訳モデルを提案する。このモデルでは、従来は変換部で解くべき問題のできるだけ多くを原言語と目的言語の換言処理の問題と捉え直し、変換部では文字通りの「言語変換」すなわち原言語表現から (複数の) 目的言語表現への写像 (候補列挙) のみを行なう。

本モデルはちょうど、我々が翻訳対象言語に不慣れな場合に行なう訳出作業に似ている。例えば、中国語の知識がほとんどない状況で中日翻訳を試みる場合、一旦は辞書で (不自然でも構わずに) 日本語に置き換えてから、日本語単独で考え、より自然な日本語に言い換える。また日中翻訳の場合は、当初は日本語入力表現に対して辞書を引くが、日中辞書に項目記載のない場合もあり、その場合は別の日本語表現に一旦言い換えて辞書を引くという作業を、我々はごく当然のように行なっている。以上の観察から、二言語知識を豊富に持つのが理想的ではあるが、仮に二言語に関係する翻訳知識がある程度不足していても、原言語並びに目的言語で換言処理を十分に行なうことができれば、ある程度の機械翻訳は可能であると考えた。

以上の基本方針に基づいて、現在我々は中日翻訳を対象にして音声翻訳システム (SANDGLASS) の構築を進めている。本稿では、本提案機構の基本設計を述べ、翻訳方式の議論を行なう。

2　翻訳機構

図1に SANDGLASS の翻訳機構を示す。SANDGLASS は音声認識結果を受け取り、まず原言語換言部で換言処理を行なう。その結果は変換制御部を通して変換部に渡され、目的言語に変換される。変換が成功した場合すなわち目的言語が得られた場合、制御部は目的言語換言部にその結果を渡す。すべての原言語換言結果が変換に失敗した場合、制御部はどこが失敗したのか[1]についての情報を原言語換言部に返し、同時に異なる換言結果を要求する (図2右)。

また、SANDGLASS では換言因子 [Yam01] を言語表現とは独立に保存する。換言因子とは換言の際に作用するパラメータのことである。我々は、厳密な意味で同義となる換言は不可能であるという立場をとる。これに従えば、換言前後の表現には何らかの差異が生ずることになり、これを我々は換言因子と呼んでいる。因子となり得るものの多くは言語外情報 (発話された状況：例えばホテルでの旅行会話、話者が女性) や話者の感情、ある特定の表現に対する強調などの情報で、これらのうちいずれかの特徴があり、かつこれらの特徴抽出に成功した場合に換言因子として保存される。換言因子は音声認識部、原言語換言部、システム外のいずれからも得られる可能性がある。

[1] もしくはどの表現が変換可能なのか

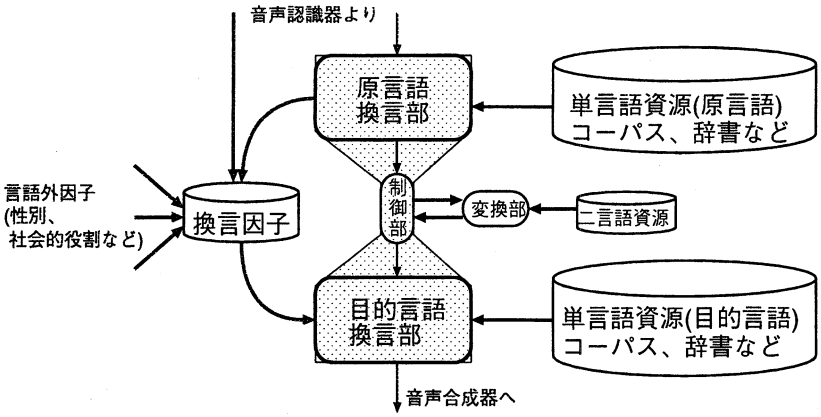

図 1: SANDGLASS: 概要

換言因子はいずれも言語に非依存であるため、これらの情報そのものに対して原言語から目的言語に「変換」する必要はない。すなわち、変換部がこれら情報を入力する必要はないと考え、本機構では変換部負荷低減の観点からこれらの情報は変換部を経由しない[2]。

最後に、目的言語換言部において、制御部から渡された複数の仮説に対して、目的言語の単言語コーパスや辞書、シソーラスなどを利用して最適な単一の目的言語表現を選択する。同時に、換言因子も入力し、目的言語において各因子を反映させた換言を行ない、その結果を音声合成器に渡す。換言因子が変換部をバイパスし、原言語換言で得られた因子を目的言語で活用することで、変換部に一切の負担をかけず、また二言語資源を一切増大させずに、原言語で持つ細かい表現の差異を目的言語において復元することを可能にする。原言語で換言因子であっても目的言語では表現の差異が生じず、換言因子とならない場合がある (性別による表現の差異や敬語など) が、このような場合は目的言語換言部の判断によってその因子を無視する。

2.1 原言語換言部

SANDGLASS において原言語で換言処理を行なう目的は以下に示すように 3 分類できる。

1. 原言語の持つ曖昧性解消
   語義の特定、並列構文の構造特定など。これらの処理はどのような機械翻訳モデルを考えた場合にも何らかの形で解決する必要のある問題であり、機械翻訳共通の問題と言い換えることができる。これについては 3 節で議論する。
2. 音声言語への対応
   音声認識誤りへの対応、話し言葉特有の言い淀み、言い誤りへの対応など。話し言葉にしか出現しない表現を一般的な表現に換言する処理も、これに含まれる。これらはすべて、テキストに対する機械翻訳では存在しない問題であり、音声翻訳に帰属する問題と捉えることができる。
3. 変換部への対応
   入力発話の翻訳単位への分割、再換言 (変換失敗後の変換可能な表現への換言) など。これらは本機構が変換部を可能な限り小規模化したことに起因する処理であり、すなわちSANDGLASS独自の問題である。

なお、中国語の換言処理については [Zha01] で、換言による曖昧性解消については [Yam01] で議論する。

2.2 変換部

変換部では、原言語表現に対して目的言語への写像を試みる。従来の機械翻訳モデルにおける変換処理との重要な差異は、本機構における変換部が以下の 2 項目の義務を共に負わない点にある。

仮説生成義務 変換部が入力された表現に対して必ず仮説を生成する義務

仮説選択義務 変換部が処理中に生起する複数の仮説に対して取捨選択や尤度付与を行なう義務

すなわち、前者により、変換部は一部の入力表現に対して変換に失敗することを機構として認める (代わりに原言語換言部が変換を成功させる表現に換言する義務を負う)。また、後者により、変換部は変換結果を一つに絞り込むことを期待されておらず、また変換時に

[2] ある表現の強調には言語表現が関係するが、「強調という情報」を変換する必要はない。すなわち原言語で強調されていた表現が目的言語でどれにあたるかを把握し、目的言語換言部で強調表現に換言すればよい。

表 1: 「都饭店预约处」という入力に対する変換結果

構文構造	目的言語への変換結果
[⟨IP⟩,[⟨NP⟩,[⟨NP⟩,[NR, 都饭店]],[NN, 预约]],[⟨VP⟩,[VV, 处]]]	/ 都ホテル / 予約 / は / 処理する /
[⟨IP⟩,[⟨NP⟩,[NR, 都饭店]],[⟨VP⟩,[VV, 预约],[⟨VP⟩,[VV, 处]]]]	/ 都ホテル / は / 処理する / ことを / 予約する /
[⟨IP⟩,[⟨NP⟩,[NR, 都饭店]],[⟨VP⟩,[VV, 预约],[⟨NP⟩,[NN, 处]]]]	/ 都ホテル / は / 場所 / を / 予約する /
[⟨IP⟩,[⟨NP⟩,[⟨NP⟩,[⟨NP⟩,[NR, 都饭店]],[NN, 预约]],[NN, 处]]]	/ 都ホテル / 予約 / 場所 /
[⟨IP⟩,[⟨NP⟩,[NR, 都饭店],[NN, 预约处]]]	/ 都ホテル / 予約係 /

仮説が増えることも認める (代わりに、原言語換言部が原言語の曖昧性を減らす義務を、目的言語換言部が変換結果候補の中から選択する義務を負う)。従来変換部に課せられていたこれらの義務は、すべて二言語知識と二言語処理の増大を招くため、英語が翻訳対象言語でない場合にこれらの負荷を変換部に与えるのは不適当であると主張する。

結局、SANDGLASS における変換部は、可能な場合にのみ目的言語候補を列挙するにとどめる。一般に、従来の機械翻訳モデルでは変換部に対して変換精度 (accuracy) と変換可能表現の網羅性 (coverage) が同時に求められていたため、どのような変換機構であってもこれを両立させるために構造的な困難性を抱えており、これが従来の典型的な機械翻訳機構の本質的な問題の一つと我々は考える。一方、本翻訳機構の変換部では、両者の要件のうち重要なのは変換精度のみであり、網羅性を上げるために変換精度を犠牲にする必要が一切ない(代わりに原言語換言部が網羅性向上を担う)。もちろん変換部においても網羅性が高いほど望ましいが、網羅性の高さは変換精度の高い場合にのみ意味を持つ。翻訳機構全体としては、以上の条件を満たす変換部であればどのような機構であってもよく、また換言処理の機構とも独立である。

ところで、表層的な情報で (特に活用語に対して) 多くの品詞を推測できる日本語とは異なり、孤立語である中国語はそのような標識を持たない。また、内容語と機能語がかなり明確に分離している英語とも異なり、中国語の機能語、特に介詞 (前置詞) の多くは内容語 (特に動詞) からの転成であるため多品詞性が顕著であり、その結果名詞／動詞や介詞 (前置詞) ／動詞など、複数の品詞を持つ語が英語よりも多い。このため形態素解析のみで品詞を決定するのは賢明ではないと考え、多品詞性を保持したまま構文解析ができる MSLR[3] を採用した。

変換部では、MSLR を用いて形態素解析と構文解析を同時に行なった後、文字列照合処理によって目的言語に変換する。例えば、「都饭店预约处」という例文に対しては、表 1に示す 5 候補が出力される。これは、「预约」が「予約／予約する」、「处」が「場所／処理する」、「预约处」が「预约＋处／预约处 (予約係)」という曖昧性を持つためである。前述したように、変換部においてはこれらの 5 候補の品詞／構文／意味上の曖昧性解消を行なわず、5 候補すべてを目的言語換言部に渡す。

[3] http://tanaka-www.cs.titech.ac.jp/pub/mslr/

表 2: 「打开」と「打碎」に対する訳語

中国語	↦ 換言結果	日本語訳語
打开书本	↦ 翻开书本	(本を) 開く
打开电视	↦ 开开电视	(テレビを) つける
打开局面	↦ 展开局面	(局面を) 打開する
玻璃杯打碎		(グラスが) 割れる
		(グラスが) 壊れる
		(グラスが) 砕ける

対訳辞書には各原言語の語句に対し一般に複数の訳語が付与されている。このような場合は、複数の訳語を併記したまま、目的言語換言部に情報を渡す。例えば、「今天要住 你们酒店(今日そちらのホテルに泊まります)」の「住」は「住む／宿泊する」という複数の訳語を持つ。同様に「酒店」は「ホテル／酒屋」という複数の訳語を持つ。この結果、変換部において両方の訳を併記したまま、「/ 今日 / の / ホテル: 酒屋 / に / 住む: 宿泊する / ＋ます /」というような形で目的言語換言部に渡される。

2.3 目的言語換言部

変換終了後、目的言語において換言処理を行なう目的は (a) 仮説の最終的な曖昧性解消 (b) 換言因子を反映したより「自然な」言語表現への変換の 2 点である。例えば、前述の例文であれば、変換部より渡された 5 候補に対して様々な尤度 (例えば別途用意した単言語コーパスにおける出現確率) により比較を行ない、「都ホテル予約係 (です)」が選択され、出力される。

3 訳語選択に対する方針

機械翻訳における曖昧性解消問題のうち、大きな問題の一つは訳語選択問題である。これは従来、一般

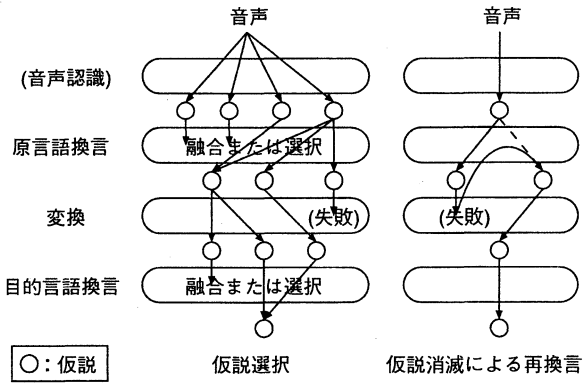

図 2: SANDGLASS における候補選択

的には変換部で解消されるべき問題と考えられているが、SANDGLASS では (1) 原言語での語義決定 (2) 目的言語での訳語決定、の二つの問題と分離して捉える。

例えば、表 2に示すように、中国語の「打开」は少なくとも「開く／つける／打開する」の 3 通りに、「打碎」も少なくとも「割れる／壊れる／砕ける」の 3 通りに訳すことが可能である。どちらの場合も、これらのうちどの訳が適切かは従来変換部で決定されていた。しかし、「打开」は中国語で三つの異なる語義を持ち、「打碎」は中国語の同一の語義が日本語で 3 通りに訳される。このような場合、SANDGLASS は

1. 「打开」は原言語で語義決定されるべきと考え、原言語換言部で語義決定と共により明確な (=多義性の低い) 語に換言する
2. 「打碎」は原言語では単一の語義が変換時に複数の目的言語の語に写像されてはじめて生じる問題のため、目的言語で解消すべきと考え、目的言語換言部でより適切な語を選択する

という方策を採用する。

ただし、原言語換言部において語義決定できなかった場合、あるいは語義決定は可能でも適切な換言ができなかった場合は、原言語において語義決定のための換言処理が行なえないが、このような場合にも目的言語において訳語決定することを試みる。原言語における語義決定の機構と目的言語における訳語決定の機構は処理内容も使用する言語資源も完全に独立であるため、この結果 SANDGLASS は独立した二つの語義選択機構を持つと考えることもでき、この意味においてより頑健なモデルとなっている。

図 2に、SANDGLASS における候補選択の概要を示す。従来の機械翻訳は語義選択、構文的曖昧性、訳語選択などほとんどの曖昧性を変換部において同時に解消する枠組みが多いが、SANDGLASS では、変換部において曖昧性の解消は行なわない。すなわち、SANDGLASS における曖昧性解消は、原言語換言部と目的言語換言部が分担して行なう。

4 まとめ

中日翻訳など英語以外の言語間で音声翻訳機構を構築するための基本提案を行なった。我々は、これら言語間の機械翻訳では、可能な限り二言語に直接関わる翻訳知識 (二言語知識) を低減し、原言語及び目的言語の単言語処理に基軸を置くことで機械翻訳機構を構成する本機構が現実的な全体設計であると主張する。ここでは従来では変換部が負っていた仮説生成義務は原言語換言部が負い、一方で変換部の仮説選択義務は原言語換言部と目的言語換言部の両者が負う。

音声翻訳システム SANDGLASS は本稿で述べた提案内容に基づいて現在実装を行なっているが、どこまで二言語知識の縮小が可能かについて現時点で検討が十分ではない。このため、特にこの点について今後検討を行なっていく。換言処理については、事例の収集 [Shi01]、現象の分析 [Zha01]、モデルの検討をそれぞれ原言語と目的言語に対して同時に進めており、これらについては別途報告する。

参考文献

[Shi01] 白井諭, 山本和英 : 換言事例の収集ー日英基本構文を対象としてー, 年次大会講演論文集, 第 7 回, P1-12, 言語処理学会 (2001).

[Yam01] 山本和英 : 換言処理の現状と課題, 年次大会ワークショップ論文集, 第 7 回, 言語処理学会 (2001).

[Zha01] 張玉潔, 山本和英, 坂本仁 : 中日音声翻訳のための中国語換言処理の分析, 年次大会講演論文集, 第 7 回, P2-14, 言語処理学会 (2001).