2005－SLP－58（2）
2005／10／21

中国語形態素コーパスと言語モデルの評価

―パラレルコーパスにおける日英両言語との比較

胡　新輝，山本　博史，菊井　玄一郎

ATR 音声言語コミュニケーション研究所
〒619-0288 京都府相楽郡精華町光台２－２－２
Email{xinhui.hu,hirofumi.yamamoto,genichiro.kikui}@atr.jp

あらまし　音声言語処理における統計的手法の有効性は、多く従来研究により示されている。統計的な音声言語処理は、日本語のみならず、様々な言語において有効な手法であるが、言語ごとに独自の特徴があるため、同じ方式で構築したモデルでも、対象とする言語により性能差が生じる。その差を調査分析し、定量的に評価することが、各対象言語における性能を改善する上で役立つと考えられる。
本報告では、ATR で整備されている旅行会話を中心にした中国語形態素コーパスを対象とし、その品質や音声認識における言語モデルの性能を対訳関係にある日本語や英語のコーパスと比較することで評価を行う。評価項目としては、各言語の語彙数、延べ語数の統計量、モデルの項目数とその頻度分布、統計言語モデルのテストセット Perplexity、および、音声認識の結果を用いた。その分析と比較を通じて、中国語コーパス及びモデルの言語的な特徴を把握し、音声認識性能を劣化させる問題点の解明を目指す。

Evaluation of Chinese Morphological Corpus and Language Models

― Comparison with Japanese and English by a Parallel Corpus

Xinhui Hu, Hirofumi Yamamoto, Genichiro Kikui

ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288, Japan
Email{xinhui.hu, hirofumi.yamamoto, genichiro.kikui}@atr.jp

Abstract： Statistical methods are proved to be effective in speech and natural language processing for various languages. Since there exist specialties with the language being processed, there are many differences among the language models of different languages even they are built in a same way. To investigate and analyze these differences are useful for improving the performance of the language processing.

In this paper, with the help of a parallel morphological corpora composed by Japanese, English and Chinese, we will evaluate the Chinese morphological corpus and language models by comparing the distribution of utterance length, parameter items of language models. We will also analyze the characteristics of the perplexities for different language models, and find the most suitable one for Chinese speech recognition. Using these analysis and experiments, we will investigate the reason that impede the Chinese speech recognition improvement.

1. まえがき

ATR では、音声翻訳システムを開発するために、旅行会話を中心とした、日本語、英語、および中国語の形態素コーパスを開発している。本コーパスは、日本語、英語、

中国語すべてが対訳関係になっており、3ヶ国語のパラレル形態素コーパスである。現在、これらのコーパスに基づいた言語モデルや翻訳モデルを、音声認識及び自動翻訳の研究・開発に利用している。しかし、言語ごとの特徴の差に起因して、各システム性能もまた異なったものとなっている。我々は、先の報告[1]で２０万文の中国語形態素コーパスの整備に関して、形態素解析仕様書の設計、言語モデルの評価及び音声認識の実験結果を示した。そこで、ある程度の知見は得られたものの、日本語と比べて音声認識性能等、まだ不十分な点があった。本報告では、前回の報告時から、更に３０万文のコーパスの追加し、形態素解析仕様に対しても部分的な変更を行った。本報告は、今回更新した中国語コーパスを対象に、その形態素精度及び言語モデルの性能を評価することを目的とし、そのために、日本語、英語のパラレルコーパス、およびそれを用いて作成された言語モデルと比較を行うことによって、中国語のコーパスと言語モデルを評価する。

以下、第２章では、ATR の 3 ヶ国語のパラレル形態素コーパス概要に関して述べるとともに、中国語コーパスについての説明を行う。ここでは中国語の形態素の定義方法について説明し、その後、日英各言語における形態素の長さの分布、語彙サイズなどを比較分析する。第３章では、言語ごとに４種類の言語モデルを作成し、その性能を比較することによって、日本語、英語、中国語各言語の違いが言語モデルの性能に与える影響を検討するとともに、テストセット perplexity によって中国語の言語モデルの性能を評価する。第４章では、それらのモデルを用いた音声認識実験結果および、認識誤りの傾向に関する考察を行う。第５章はむすびであり、現状の中国語コーパス及びモデルに存在している問題と対策について議論する。

2．対象コーパス

2.1 ATR パラレルコーパス

ATR では、頑健な音声翻訳システムを構築するために、多数の言語表現を含むコーパスを収集し、整備している。これらのコーパスは、元々、日英音声翻訳のために、さまざまな収集方法で、かつ複数のドメインを対象とした日本語や英語の対話データであった。データの収集方法とドメインによって、以下のようなデータセットが整備されている。

SLDB（Spoken Language Database）は人間の通訳者を介した対話であり、対話者及び通訳者の発話を収録したものである。

BTEC（Basic Travel Expressions Corpus）は旅行中の様々な場面で用いられると考えられる表現を書き出して対訳を付与したものである。

MAD(Machine Translation Aided Dialogue)は ATR の音声翻訳システムを介して日英の話者に対話を行わせ[1]、その発話を収集したものである。

現在、これらのデータのうち、一部の日本語に対して英語とともに、中国語に翻訳がなされている。従って、本研究が対象としている中国語データは、日本語、英語、中国語すべてが対訳関係になっており、三ヶ国語のパラレルコーパスになっている。

2.2 中国語形態素コーパス

我々は、中国語コーパス整備作業をするために、北京大学計算語言研究所（PKU）の「現代漢語語料庫加工規範―詞語切分与詞性表注」という形態素解析仕様書をベースとし、対話を対象とする音声翻訳を目的として、中国語の形態素解析仕様書を定めた[2]。この仕様書に基づいて、自動処理とその結果に対する人手の修正を行うことによって、表 1 に表された 54 万文の中国語形態素コーパスを整備した。その中に、ランダムに 3.1 万文を選び、テストデータとして定めた。

表 1．中国語及びその対応する日本語、英語のコーパス[3]

	発話数	中国語延べ語数	日本語延べ語数	英語延べ語数
学習セット	510K	３．５M	４．３M	３．８M
テストセット	31K	204.3K	２６７K	２７２K

2．2．1 形態素の定義と発話の長さ

日本語と同じように、中国語の文章は英語のような単語間のスペースが存在していない。従って、自然言語処理の前処理として、形態素解析、すなわち漢字列である文章を形態素単位に区切り、さらに品詞を付与ことが必要となってくる。しかし、中国語では、いくつかの形態素の定義が存在する。PKU はその中で最も一般的な定義であり、他の定義と区別するために「セグメンテーション単位」という名称を用いている。PKU の仕様では、それを用いて、いかに一つの文を幾つかの短い文字列に区切るのかを決めている。我々の中国語の形態素解析仕様では、原則としてその「セグメンテーション単位」を中国語の形態素と見なす。我々の仕様で定義されているセグメンテーション方法は、基本的に PKU の仕様書に準拠している。その理由としては、PKU の仕様は、中国の言語学学者と自然言語処理の専門者達が長年を経て制定した標準であること。それに基づいて整備された巨大な人民日報のコーパスが開発済みであること。多くの中国語自然処理の分野で採用されていることがあげられる。

---

[1] 最新版の MAD データの一部は、ATR の日中音声翻訳システムを介した日中の話者の対話により収集したものがある。

[2] 参考文献「１」に定義した品詞表に、更に擬声詞（o）を追加した。

[3] ここに提示している日本語と英語は、中国語に訳した部分のみである。また、テストセットは、中国語に対応する部分以外のものを含む。

[4] ここに提示している日本語と英語は、中国語に訳した部分のみである。また、テストセットは、中国語に対応する部分以外のものを含む。

一方 ATR において、日本語と英語も同じように、音声言語処理のために、それぞれの形態素解析仕様が定義されており、それらに従って、コーパスの整備を進められている。ここで各言語に対する分析は、これらの整備済みのコーパスに対して行う。

表１に示すように、対訳関係にある日本語（JP）、英語（EN）及び中国語（CH）の形態素コーパスデータに含まれる延べ語数及び語彙サイズは異なっている。その中で、日本語の延べ語数が最も多く、英語がその次で、中国語が最も少ない。一方、語彙サイズは、中国語が最も多く（47,300）、次いで、日本語（45,500）、英語（32,900）の順となっている。次に示す例文は、BTEC の学習データから取り出した各言語の文例である。[]中はその文に含まれる形態素数である[1]。

例文：

（１）CH:没关系/ [1]
JP:大丈夫/ です/ [2]
EN:i’m/ Okey/[2]

（２）CH:把/ 钥匙/ 忘/ 在/ 房间/ 里/ 了/ [7]
JP: 部屋/ に/ 鍵/ を/ 忘れ/ まし/ た/ [7]
EN: i/ locked/ myself/ out/ [4]

（３）CH: 没有/ 住院/ 的/ 必要/ [4]
JP：入院/ する/ 必要/ は/ あり/ ません/ [6]
EN: you/ don’t/ need/ to/ be/ hospitalized [6]

（４）CH: 有/ 会/ 说/ 日语/ 的/ 人/ 吗/ [7]
JP:日本語/ を/ 話せる/ 人/ は/ い/ ます/ か/ [8]
EN: is/ there/ anyone/ here/ who/ can/ speak/ japanese/ [8]

日本語では、「は、を、が、か」などの助詞が存在しており、種類も多く、出現頻度も高い。一方、中国語には日本語の助詞に相当する役割をする形態素が少ない。中国語では、「吗」のような疑問文の結尾に使われる語気助詞、動作の完了を表す語気助詞「了」、所属関係を表す構造助詞「的」が使われているが、種類が比較的少ない。

図１は、各言語の５０万文の学習データから得られた発話の長さ（一文に含まれる形態素の数）の分布である。図に示すように、一発話あたりの形態素数は中国語が一番少なく（平均 6.95）、その次は、英語で（7.74）、日本語は一番多い（8.60）ことが分かった。

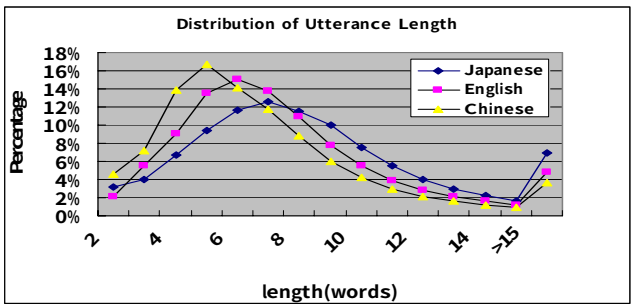

図１．各言語の発話の長さの分布（単位：単語数）

[1] この数字は、文末の句読点を含んでいない。

2.2.2 語彙サイズ

表１に示したように、中国語では、延べ語数が少ないにもかかわらず、語彙サイズは大きい。その原因は、以下の点が考えられる。

- 中国語では、同一表層でありながら、複数の品詞を持つ単語が多い。
- 同じ原文の形態素が異なる中国語形態素に訳されている。これは、特に固有名詞の場合に多く発生している。例えば、地名“アドレト”は、“阿德莱得”と“阿德莱德”に訳している。

多くの語が接辞的に用いられるため、複合語が生成されやすい。例えば、“安全”から、“**安全带，安全帽，安全门，安全套，安全箱，安全性，安全椅**”と複合語が生成される。

この中で、最初の項目は、品詞付けの誤りの大きな原因ともなり、人手による修正作業量の増大の原因ともなる。最後の項目は中国語の特徴であり、詳細な分析を行う必要がある。

これに対し、日本語の辞書サイズが多いのは、以下のことが考えられる。用言の各活用型に対し、活用形（未然形・連用形・終止形・連体形・仮定形・命令形など）が 16 個設けられている。これらの活用形はたとえ表層形が同じであっても辞書項目としては別に登録されている。例えば、「歩く」の活用形「歩け」は、辞書に対し、いかの五つが登録されている。

(1) 歩け|アルケ|歩く|本動詞|五段カ|ば|| [歩け]
(2) 歩け|アルケ|歩く|本動詞|五段カ|命令|| [歩け]
(3) 歩け|アルケ|歩ける|本動詞|一段|ない|| [歩け]
(4) 歩け|アルケ|歩ける|本動詞|一段|連用|| [歩け]

3．言語モデルの評価

3.1 評価用モデルの説明

音声翻訳システムの研究・開発に必要とされるデータ、特に話し言葉のデータ収集は、非常に困難である。このため、システム構築では、限られた量のデータを使って、言語モデルの訓練を行うことが多い。この場合、通常の単語 N-gram ではデータスパースの問題が起こり易く、これを解決するための、幾つか手法が提案されている。代表的なものは、クラス N-gram である。単語間の遷移確率は単語に属するクラス間の遷移確率として近似される。ATR では、このクラスベース方法を拡張し、マルチクラス N-gram、及びマルチクラス複合N-gram を提案している[１]。これらのモデルはすでに日英音声翻訳システムにおいて、その有効性が確認されている。今回、さらに中国語における有効性を確認するために、以下の四つのモデルによる評価を行う。

（１）単語 2-gram (w2)
（２）単語 3-gram (w3)
（３）マルチクラス 2-gram (MC2)
（４）マルチクラス複合 2-gram (MCC2)

ここでの評価内容は、モデルのサイズ、テストセット Perplexity、各言語の文の平均エントロピーなどである。

3.2 評価実験

以下、ことわりのない限り、学習データは、表１の５０万文の学習データで、テストセットは、表１に示したテストセットから選んだ 1524 文である。

3.2.1 各言語のモデルの比較

各言語の学習データから同じ対訳の１６万文を用いて、単語 3-gram を作った（Cutoff=5, Smoothing＝Good-Turing)。表２は、それらの 3-gram のパラメータ数を表す。また、図２は、それらの中に出現した 2-gram 出現頻度の割合を示す。これにより、学習データに含まれる形態素数は中国語が一番少なく、モデルのパラメータ数は、中国語が一番多いということが分かる。

次に、テストセット文の総形態素数と平均エントロピーを、表３に示す。文あたりの平均エントロピーは中国語が一番大きく、日本語が一番小さい。つまり、本中国語コーパスでは、文を推定するために必要な情報量は日本語、英語より多いことがわかる。

表２．各言語の 3-gram におけるパラメータの比較

	中国語	日本語	英語
Words	1, 043, 960	1, 286, 790	1, 140, 860
Parameters	177, 668	166, 214	153, 734

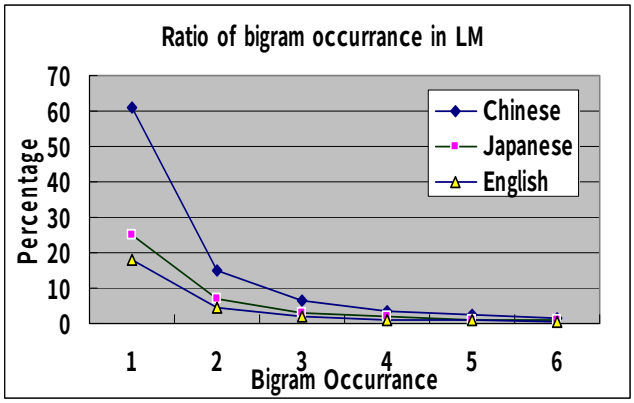

図 2. 各言語のモデルの bigram エントリの出現回数の比率

表３．各言語のテスト文の平均エントロピー

	中国語	日本語	英語
テストセットの語彙サイズ	10, 030	12, 344	10, 840
文の平均エントロピー	294. 58	165. 80	202. 92

また、他の言語と比べて、中国語には、単に語彙サイズが大きいだけではなく、頻度の低い 1-gram、2-gram 及び 3-gram のエントリも多い。例えば、頻度が 1 の 2-gram エントリが 2-gram エントリ全体の６０％を超えている。これらの 2-gram エントリの確率は、平滑化によって推定せざるを得ず、全体的に精度の低下を起こす。つまり、中国語の場合に、データスパースの問題が、他の言語より大きいと考えられる。

3.2.2 学習量と Perplexity の関係

テストセットを固定で、学習データ量を変化させた場合の 3-gram を作成し、それらのテストセット Perplexity を求めた。その結果は、図 3 に示されている。学習データの増大により、何れの言語も Perplexity の値が小さくなる傾向がある。但し、学習データがある量に至れば、その減少傾向が小さくなり、Perplexity の減少が飽和する。

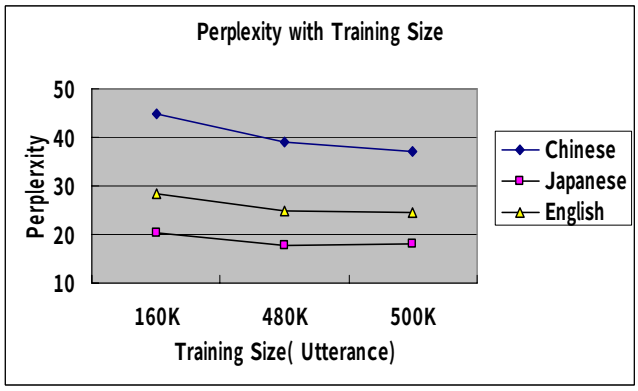

図 3. Perplexity と学習量の関係

3.2.3 中国語の各モデルの比較

図 4 は各言語モデルに対する中国語のテストセットの Perplexity を表す曲線である。クラスの数の増大に伴って、マルチクラス 2-gram とマルチクラス複合 2-gram の Perplexity が低くなる傾向を示している。また、マルチクラス複合 2-gram の Perplexity は、マルチクラス 2-gram より低い。これは、[3]において日本語と英語のモデルで得られた結果と同じである。単純に perplexity の値からモデルを評価すれば、単語 3-gram,マルチクラス複合 2-gram、単語 2-gram、マルククラス 2-gram の順となる。但し、モデルの優劣は、他の様々な要因を考慮しなければならない。

一方、各言語モデルのサイズは、表３に示すように、マルチクラス 2-gram とマルチクラス複合 2-gram のパラメータ数が単語 2-gram の約１4―16%に留まっており、コンパクトで高性能のモデルであることを示している。

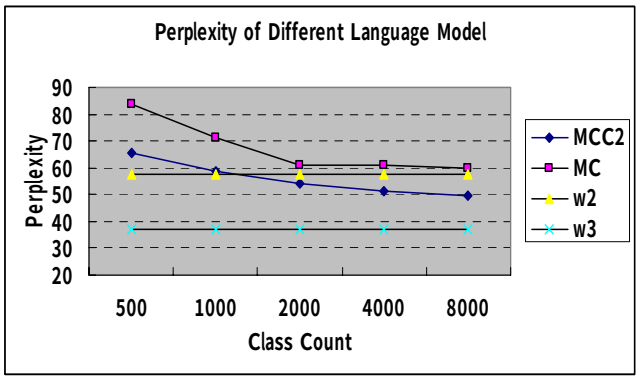

図 4. 中国語の各言語モデルの Perplexity

表４．中国語各モデルのパラメータ数[1]

モデルの種類	パラメータ数
単語 2-gram	489，370
単語 3-gram	1，636，634
マルチクラス２-gram	82，232
マルチクラス複合 2-gram	68，621

４.音声認識実験

上述のコーパス及び言語モデルは、音声翻訳システムの構築を目的として整備されている。従って、その品質は、実際の音声翻訳システムにおいて確認する必要がある。例えば、Perplexity は言語モデルの性能を表わす良い尺度であるが、必ずしも音声認識における認識率に直結することは限らない。そこで、前述の Perplexity 評価で用いた言語モデルとその評価データの 510 文を用いて、連続単語音声認識による評価を行った。

音響モデルは、540 名（男女毎に 270 名）の話者が録音した２１万発話の中国語音声データベースによって訓練される音素環境依存 HMNET である（ML-SSS 法、1200 状態、５混合、性別依存)。このデータベースに、北京、上海、広東及び台湾アクセントを含む中国語音声データである。

音響モデルの分析条件は、以下のようになる：サンプリング周波数は１６KHz、特徴量は MFCC,ΔMFCC、Δ対数パワーである。

4.1 多言語の比較

表１に示している各言語の学習データから約１６万文を用いて、単語 2-gram を作り、音声認識実験を行った。その認識率を、図 5 に示す。

ここで、WordID は、単語出現形、品詞及び他の情報（日本語の場合は、活用など）を考慮しその全てが一致した場合のみを正解として扱っている。これに対し、Surface Word は、表層形のみを考慮した場合である。この図から、日本語と英語は、この両者の認識率の差は１％程度であるが、中国語は２％を超えていることが分かる。コーパスに対する品詞付けの難易度、および精度が、この差の最大の原因であると考えられる。

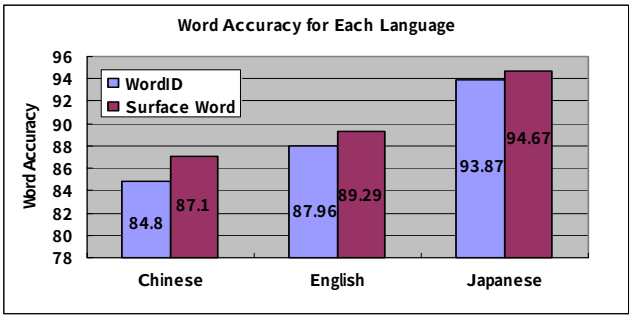

図 5. 各言語の単語 2-gram の単語認識率

[1] マルチクラス 2-gram 及びマルチクラス複合 2-gram に、クラス数＝2000.

4.2 中国語の各言語モデルの比較

図 6 は、各種のモデルを使った中国語の音声認識の結果である。図に示されるように、マルチクラス複合語 2-gram が、最も高い性能を示しており、日本語、英語と同様の傾向である。マルチクラス 2-gram は Perplexity では単語 2-gram よりも高いが、認識率では勝っている。この原因は、マルチクラス 2-gram の方がデータスパースの問題して頑健であると考えられる。

また、品詞を考慮した場合の単語認識率は、表層形のみを考慮した場合の単語認識率と比べ、２％以上の差があることが、すべてのモデルにたいして言える。

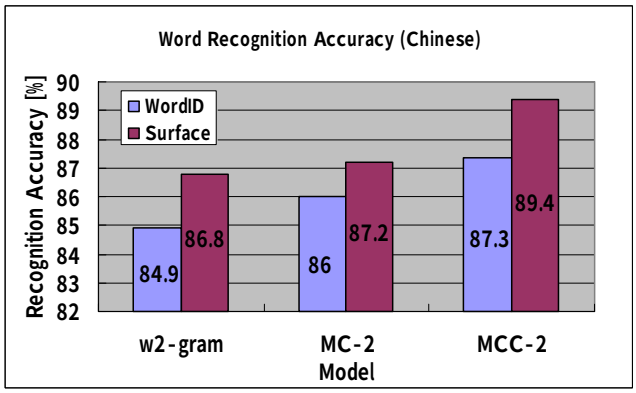

図 6. 中国語における各モデルの単語認識率

4.3 言語モデルによる認識誤りの分析

音声認識における認識誤りは、音響モデルに起因するものと、言語モデルに起因するものがある。ここでは、言語モデルに起因する認識誤りについて調べるため、上述の正解音素列を既知とした場合、すなわち、音響モデルに誤りがないと仮定した場合についての分析を行った。ここでは、あたえられた正解音素列を生成するような形態素列のうち、言語尤度が最も高いものを音響モデルに誤りがない場合の認識結果としている。そして、この結果に対する誤りを分析することにより、言語モデル単独の性能をおよび問題点を把握する。

この目的のために、誤りを三種類のカテゴリに分類した。

（１） セグメンテーションエラー（２６個）

例[2]：不/同/—>不同/，这个/—>这/个/,不/是/—>不是/

（２） 品詞付けエラー（４５個）

例：的/de—>的/y,给/p—>给/v,中心/n->中心/ns,去/v—>去/vt,在/p—>在/v,对/a—>对/p

（３） 未知語、同音異義語、同声調語などのエラー（４１個）

例：十足—>是/租，那个/哪个，除了—>出/了，吉他—>吉它，格拉斯歌—>格拉斯格，黒胡椒—>嘿/胡椒

セグメンテーションエラーの約半分は学習コーパス自体の誤りが原因である。例えば、“不同—>不/同”である。また一部は、セグメンテーション自体の曖昧性によって

[2] →の左は、リファレンス結果で、右は、認識結果である

発生したものもある。例えば、“不/是”と”不是“両方共存在し得、コンテキストによって判断しなければならない。

品詞エラーでは、少数のパターンに集中している。

- 普通名詞（n）と処所詞（ns）.

これは、仕様書の定義の曖昧性が原因であると考えられる。処所詞の定義を明確する事により、改善をはかることができると考えられる。

- 前置詞（p）

品詞付けの誤りの中で一番頻度が高いものである。動詞と混同するケースが多い。大多数の前置詞は、動詞から派生したもので、使用頻度も高い。一般的に、前置詞は名詞と一緒に用いて場所、時間、原因などを表し、文中にある動詞にかかる。人間は関連名詞や動詞が存在しているかどうかによって、前置詞を判断することができる。しかし、前置詞と名詞或いは動詞の間の間隔は自由であるため、N-gram 方式では限界があると考えられ、別の対策が必要と考える。

- 傾向動詞（vt）と能願動詞（vw）

傾向動詞や能願動詞は通常、普通動詞と共に用いられる。誤りの原因の多くは、普通動詞を伴わずに用いられている場合であり、この場合はやはり N-gram 方式では限界があると考えられる。

（3）のエラーは最も複雑である。同音同義異表記語の問題はコーパス整備時点で表記を統一することで改善できる。また、同音異調語は、音響モデルを声調に対応すれば解決できる可能性があるが、同音同調異義語に対しては別の方法を考える必要がある。

５．むすび

本研究では、ATR で整備している中国語対話コーパスと言語モデルの評価のために、中国語コーパスと対訳になっている日本語や英語コーパスを用い評価を行った。評価項目は、各言語での延べ語数、語彙サイズ、発話の長さ、2-gram エントリ数とその頻度分布などの特徴を比較し、現状のコーパスとモデルにおける問題点を明らかにした。また、各言語のテストセットのエントロピーを比較することにより、中国語のデータスパースの問題が、他の言語比べて大きいということが分かった。また、中国語音声認識における種々の言語モデルの比較では、日本語や英語の場合とほぼ同じ傾向を持っており、マルチクラス複合 N-gram が最も高い認識性能を持つことが示された。また、認識エラーを分析するにより、コーパス精度の問題の他、中国語自身の特徴により起こる問題があることが判明した。例えば、前置詞、傾向動詞、能願動詞の標柱に対して、単に N-gram 方法では限界があるため、他の方法が、例えば、ルールベース方式、求められる。今後、まず、コーパスや言語モデルの 2-gram 項目の比較で得られたデータ分布により、コーパス精度の改善を図る。その後、更にそれらに対し定量的な分析を行い、データスパース問題に対する対応を行う予定である。

謝辞

本研究は、総務省からの研究委託「携帯電話などを用いた多言語の自動翻訳システム」により実施したものである。

参考文献

[1] 胡新輝、劉敏、山本博史、菊井玄一郎、「音声翻訳のための中国語コーパスの整備とその評価」　情報処理学会研究報告、2005-NL-167, 2005-SLP-56, pp47-52, May, 2005.

[2] 菊井玄一郎，竹澤寿幸，山本誠一，“対話翻訳のための音声言語コーパスの現状”,日本音響学会２００４年春

[3] Hirofumi Yamamoto, Shuntaro Isogai, Yoshinori Sagisaka,“Mutli-class composite N-gram language model”, Speech Communication, 2003, Vol.41, pp369-379.