科学研究費助成事業　　研究成果報告書

平成 ２６ 年　 ６ 月　 ３ 日現在

機関番号：１２１０２
研究種目：挑戦的萌芽研究
研究期間：2012～2013
課題番号：２４６５００６３

研究課題名（和文）分岐ｎｇｒａｍモデルによる短距離言語モデルから中距離言語モデルへの飛躍

研究課題名（英文）A leap from short range language models to middle range modeling using dependency ngrams

研究代表者
山本　幹雄（YAMAMOTO, MIKIO）

筑波大学・システム情報系・教授

研究者番号：４０２１０５６２

交付決定額（研究期間全体）：（直接経費）　　2,900,000円、（間接経費）　　870,000円

研究成果の概要（和文）：　声認識や統計的機械翻訳システム等の言語モデルとして、現在、ngram言語モデルが広く利用されているが、このモデルは隣り合った単語の連鎖の確率に基づくモデルである。完全に語彙化しているモデルであるため、局所的な単語の連鎖を精密にモデル化する。しかし、ngram言語モデルは文の構造を無視しているため、中長距離の言語的特長を捉えられない。本研究では、この問題を解決するために、ngram言語モデルに依存構造を統合した生成的依存ngram言語モデルを提案した。すべての依存構造を考慮することによって、任意の次数の依存ngramの確率をEMアルゴリズムによって推定可能とするアルゴリズムを示した。

研究成果の概要（英文）：　Statistical language models are a fundamental component of speech recognition systems, machine translation systems, and so forth.　Presently, the ngram language model is the most widely used approach.　This model focuses on sequences of neighboring lexical words, and uses the probabilities of these sequences as model parameters.　Due to the full lexicalization of the ngram language model, local features of word sequences can be well modeled.　However, an ngram language model cannot capture relatively medium or long-range features, because it regards a sentence as a flat string and ignores its structure.
In this research, we proposed a generative dependency ngram language model that integrates a generative dependency structure of a sentence into the original ngram language model.　Using an expectation-maximization (EM) algorithm, the probability of arbitrary order dependency ngrams can be estimated by considering all possible dependency structures of a sentence.

研究分野：自然言語処理

科研費の分科・細目：情報学・知能情報学

キーワード：確率的言語モデル　依存構造　機械翻訳

１．研究開始当初の背景

自然言語文の生起確率を与える統計的言語モデルは統計的機械翻訳や音声認識のような言語文を合成・出力するシステムになくてはならないコンポーネントであり、1980 年代から世界中で精力的に研究されてきた。1980 年代には、主に短距離の単語隣接関係に基づくモデル（*n*gram モデルと呼ばれる）が発展し、現在でも統計的機械翻訳などで主力モデルとして使われている。1990 年代に入ってからは、文の句構造などをベースに *n*gram よりも広い文全体の単語依存関係を利用するモデル化技術が進んだが、性能的には *n*gram モデルを凌駕するには至らなかった。2000 年代に入ってからは、文を超えたより長距離の文脈を統計的言語モデルに取り込む研究が盛んとなり、topic モデルと呼ばれる文章あるいは文書全体をモデル化する手法が進展した。topic モデルは、ある程度の翻訳性能向上には役立ったが、現在の統計的機械翻訳システムは 1 文単位の翻訳性能に依然問題があるため、画期的な改善は達成されていない。結局、翻訳性能を上げるためには、文の構造を用いた言語モデルの研究に立ち返り、1 文の自然さをモデル化する中距離の統計的言語モデルを構築することが近道であるという認識に至った。

２．研究の目的

本研究では、技術的にはほぼ完成している *n*gram 短距離モデルについて、文構造に沿って分岐することを許す中距離モデルを提案することを目的とする。本研究で計画したモデルの特徴は、文の構造的可能性をすべて考慮しながら文の確率分布を *n*gram モデルの拡張として定義する点と、文構造の特長もモデルに取り入れる点にある。本研究の理論的な研究範囲は、次の 2 つのアルゴリズムを提案・形式化することである：(1)すべての依存構造を考慮しながら任意の次数の依存 ngram 確率を用いて、文 S の確率 P(S)を計算する高速なアルゴリズム、(2)大量の文データからモデルのパラメータ推定を行うアルゴリズム。

さらに、提案した分岐 ngram モデルを統計的機械翻訳と文構造解析に応用し、性能を評価する。

３．研究の方法

短距離言語モデルの代表である *n*gram モデルは、1 文の確率を短距離の隣接 *n* 単語の条件付き確率に直線的に分解して計算する極めてシンプルなモデルである。本研究のアイデアは、直線的な分解を、分岐的な分解に変更することによりほぼ同じパラメータ数のモデルで文の構造を同時にモデル化する点にある。図 1 に基本的なアイデアを図解する。上段が従来の *n*gram モデルであり、太い線で示されている単語の関係の確率を掛け合わせて文全体 *S* の確率を得る。分岐 *n*gram モデルでは依存構造によって決まる構造に沿った 2 単語の確率を掛け合わせることによって、構造 *D* によって条件付けられた文 *S* の確率を得る。この図は日本語を念頭に置いているため、*n*gram の確率は一方向であるが（日本語は原則、文の前から後ろの単語に係る）、このアイデアを英語などのように係り方向が両方向である言語にも使えるように洗練する。似たモデルとして、依存構造上の主辞同士の依存確率を用いたモデルが知られているが、本モデルは条件部分の重なりを許す点であくまでも *n*gram モデルの拡張である点で本質的に異なる。

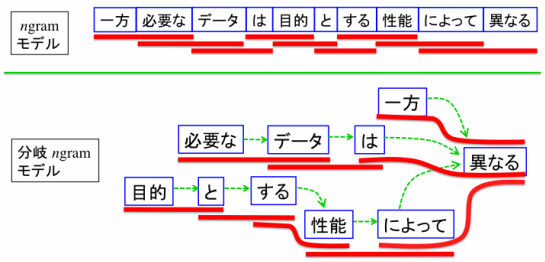


図 1：分岐する ngram のアイデア

基本的なモデルに基づいて、中距離言語モデルとしての 2 つの基本アルゴリズムを開発する。

(1) 基本アルゴリズム 1

依存構造と文の同時確率 $P(S,D)$から、周辺確率 $P(S)$を効率的に計算するアルゴリズムを開発する。確率的文脈自由文法で研究されている Inside アルゴリズムは CKY 構文解析法の三角表を利用して、入力文の長さの 3 乗のオーダの時間計算量で確率的文脈自由文法を用いた周辺確率を計算できる。これはおおよそ、本研究で考えるモデルの *n*gram を n=2(bigram)と考えたものに対応する。

しかし、*n*gram の n を 3 以上とすると、*n*gram の重なりが出て来るため Inside アルゴリズムは利用できない。そこで、三角表を n 次元に拡張する方法を考える。n=2 の場合が 2 次元であるので、n=3 の場合は 3 次元、一般には n 次元の三角表を考えればよい。3 以上の

n に対して増えた単語列を条件とした三角表を複数用いるイメージである。この方法を追求し、完成されたアルゴリズムとして定式化する。

(2) 基本アルゴリズム 2

文集合を学習データとして、依存構造上の *n*gram モデルのパラメータ（すなわち、条件付確率）を最尤推定するアルゴリズムを開発する。確率的文脈自由文法の推定アルゴリズムである Inside-Outside アルゴリズムからの類推により、*n* 次元三角表上の Inside 確率と Outside 確率を定義し、EM アルゴリズムを適用することにより構成可能である。さらに、実用的な推定手法としては *n*gram モデルのスムージング技術を導入する。最も単純な実装としては、構文解析器の出力を教師データと考え、依存構造上の *n*gram をカウントし、従来の *n*gram とまったく同じ推定手法を利用する方法である。

４．研究成果

以下のような 5 つの成果を得た。(1)依存構造と文の同時確率をモデル化する新たな分岐 ngram モデルの提案。(2) 任意の次数の依存 ngram に基づく周辺確率の高速計算手法。(3)任意の次数の依存 ngram のパラメータ推定法。(4)構文解析と(5)機械翻訳への応用による評価。それぞれの成果について以下に述べる（紙面の制限で詳しいモデル／手法を述べることができないため、詳しくは「5.主な発表論文等」の論文を参照のこと）。

(1) 分岐 ngram モデル

分岐 ngram モデルは単語列と依存構造の組に対して同時確率を与える。ngram モデルでは単語の出現確率は直前の n−1 単語の種類に依存するが、分岐 ngram モデルでは依存構造木における n−1 個の先祖の単語とその関係に依存する。依存構造を利用することにより、直前 n−1 単語に限らずに離れた位置にある単語の情報を利用できる。分岐 ngram モデルは、2 単語間の依存関係に着目した依存構造 bigram(Lee and Choi, 1998)を任意の n 単語について拡張したものといえる。分岐 ngram モデルでは単語 w は全て、依存リンクの向きに応じて左側と右側に任意個数の修飾語を持つ。依存構造木の葉に当たる部分は空の列を左右に一つずつ生成する。単語 w の左側の修飾語の列を<L></L>で囲い、右側の修飾語の列を<R></R>で囲う。これらの記号をまとめて分岐タグと呼ぶ。例として"I eat pizza with Maria."という文に対して図 2 のような依存構造を考えた時、分岐タグを付与した依存構造木が図 3 である。

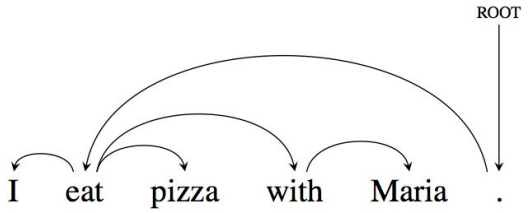


図 2　依存構造の例

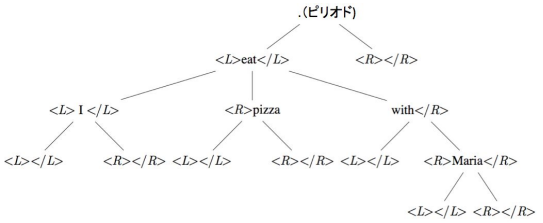


図 3　タグ付与した依存構造木

分岐 trigram による同時確率は、分岐タグを付与した依存構造木上の各ノードの確率（ただし、ルートからのパスを履歴とする条件確率）の積で近似される。確率を計算するノードとしては、ルートノードとしての'.'(ピリオド)および、<L>と<R>タグは除くが、</L>と</R>タグを含める。これによって、言語毎の左右の依存構造パターンをモデル化する。例えば、日本語では左から右に修飾する場合がほとんどであるため、</R>タグの確率は極めて低くなるはずである。また、条件となるルートからのパスには<L>と<R>タグを含めることにより、分岐の構造をモデルに取り込む。具体的には、ngram 条件付確率の条件部に、各単語が親（head 単語）の単語に対して文中で左右のどちらにあるかを加える。例えば、図 3 中の'pizza'の親ノードは'eat'であるが、'pizza'は'eat'の右から依存しているので<R>タグを'eat'とペアとして考える。全体的には、例えば、図 3 の同時確率は、分岐 trigram を使った場合、次式で求められる（条件部の単語の左側の<L>または<R>タグが分岐の方向を表している））。

P("図 3 で表される木")=
P(eat|<L>,.)×P(</L>|<L>,.)
　×P(I|<L>,eat, <L>,.)
　×P(</L>|<L>,eat, <L>,.)
　×P(pizza|<R>,eat, <L>,.)
　×P(with|<R>,eat, <L>,.)
　×P(</R>|<R>,eat, <L>,.)
　×P(</L>|<L>,I, <L>,eat)

×P(</R>|<R>,I, <L>,eat)
×P(</L>|<L>,pizza, <R>,eat)
×P(</R>|<R>,pizza, <R>,eat)
×P(</L>|<L>,with, <R>,eat)
×P(Maria|<R>,with, <R>,eat)
×P(</R>|<R>,with, <R>,eat)
×P(</L>|<L>,Maria, <R>,with)
×P(</R>|<R>,Maria, <R>,with)

最終的に、分岐 ngram モデルでは文 W の生成確率 P(W)を W に対して与えうる全ての依存構造 D との同時確率の周辺確率で得られる。

(2)分岐 ngram に基づく周辺確率の高速計算

上記(1)の最後に述べたように、P(W)を計算するためにはすべての依存構造の周辺確率を計算する必要がある。しかし、すべての依存構造を列挙することは計算量的に無理であるため、効率的な手法が必要である。本研究では、Lee and Choi(1998)の依存構造上の 2 つの単語間の関係(すなわち、bigram)をベースとした完全リンクと完全系列の考え方を、任意の次数の ngram に拡張することによって、この問題を解決した。詳しくは、発表論文を参照のこと。

(3) 分岐 ngram のパラメータ推定

確率文脈自由文法のパラメータ推定手法である Inside-Outside アルゴリズムと同等な手法を用いる。ただし、Inside 確率と Outside 確率を上記(2)で述べた、任意次数の ngram に拡張された完全リンクと完全系列を用いることによって計算できる。

Inside 確率と Outside 確率が計算されれば、その積によって、1 つの分岐 ngram の確率的カウントが計算できる。すべてのトレーニング文を用いて、各分岐 ngram の確率的カウントを合計し、分岐 ngram のパラメータ(確率)の再推定を行う。これを、収束するまで繰り返すことにより、最尤推定を行うアルゴリズムを提案した。詳しくは発表論文を参照のこと。

(2)と(3)のプログラムを実装し、perplexity(情報理論的な評価尺度)による評価を行った(値が小さいほどよい)。学習に用いたデータは、Europarl から英語、ドイツ語、スペイン語、NTCIR-8 の特許コーパスから日本語を利用した。それぞれ、37 万～48 万文程度の訓練データである。トレーニングに用いていない、テスト用の文をそれぞれ 2,000 文用意し、test-set perplexity を計測した結果が表 1 である。

表 1　Test-set perplexity

言語	bigram	trigram
英語	159	156
ドイツ語	265	261
スペイン語	159	158
日本語	88	67

日本語については、修飾関係が左から右(依存関係が右から左)のみを考慮したため、単純な構造となっており高い性能を発揮している。また、次数の高い分岐 ngram が高性能であるが、日本語以外では大きな差は見られなかった。日本語において trigram が高性能となった理由ははっきりしないが、一つの理由として学習パラメータが半分になった分、学習が頑健になったと考えられる。

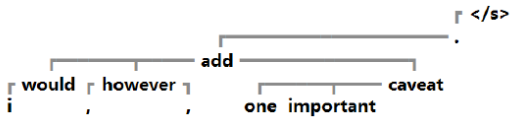
(4) 構文解析への応用による評価

2 つの応用における評価を行った。①本研究で提案した手法による教師なし学習結果を用いた文の構造解析である。②既存の依存解析器の出力した依存構造を正解データとして、分岐 ngram を学習し(Inside-Outside アルゴリズムを使わないで)、どれだけ既存の依存構造解析器に近づけるかを評価した。

①教師なし依存構造解析は、学習した分岐 ngram を用いて、入力文と依存構造の同時確率が最も高くなる依存構造を出力とした。以下に各言語の例を挙げる。本実験では、依存構造のルートノードをピリオドではなく、文の最後に付ける</s>タグとした。

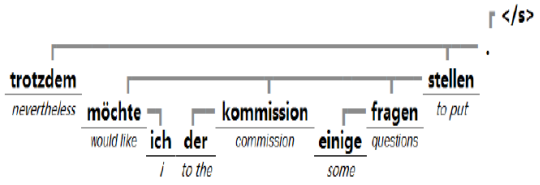
英語：

i would , however , add one important caveat .



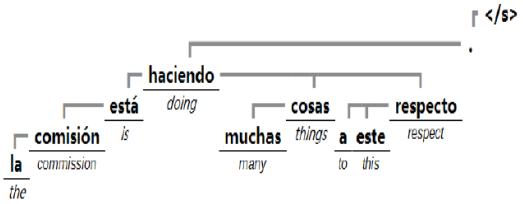
ドイツ語：

trotzdem moeochte ich der kommission einige fragen stellen .



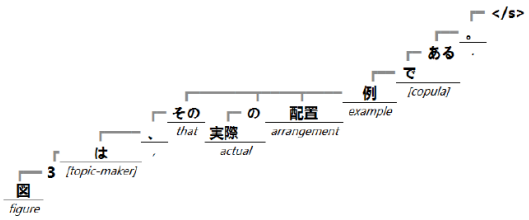
スペイン語：

"la comisi_on est_a haciendo muchas cosas a este respecto."



日本語：

「図 *3* は、その実際の配置例である。」



これらの例より、教師なしにもかかわらず、ある程度に人間に近い構造を自動的に獲得していることが分かる。

②既存の依存構造解析器のシミュレートについては、データとして NTCIR-8 の日本語特許文を Cabocha と呼ばれるフリーの解析器にかけた結果から分岐 ngram モデルのパラメータを直接推定した。その後、①の実験と同様に、入力文と依存構造の同時確率を最大とする依存構造を解析結果として出力する。学習データとは別にテスト用の文を用意し、既存の解析器の出力を正解として、分岐 ngram による解析結果を評価した。評価値は、日本語における文節を単位として、文節の係り先の文節の一致率である。100%であれば、既存解析器とまったく同じ出力であることを意味する。この一致率と訓練データ数の関係を図 4 に示す（縦軸が一致率(%)、横軸が学習データの文数)。これより、概ね 80%以上の一致率を達成できることが分かる。また、訓練データは 3000 万文で飽和しておらず、訓練データを増やせばより高い性能が出せることが分かる。

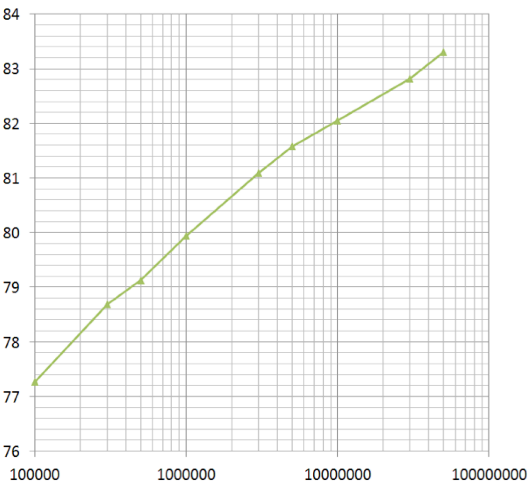


図 4　訓練データ量(横軸)と一致率(縦軸)

(5) 統計的機械翻訳への応用

統計的機械翻訳システムでは、翻訳の質を判定するために翻訳モデルと言語モデルの 2 つが用いられる。おおまかに言えば、翻訳モデルは意味の適合性を、言語モデルは文法的な適合性を判定する。言語モデルとしては、一般に従来の ngram モデルが用いられるが、今回は従来の ngram に加えて、分岐 ngram モデルのスコアを重みを付けて加えることにより、より良い翻訳ができるかどうかを評価した。

翻訳方向は英語から日本語への翻訳とし、NTCIR-8 の特許翻訳コーパス 180 万文を用いて確率モデルを推定した。従来の言語モデル、分岐 ngram ともに 5-gram モデルを用いた。スコアに加える分岐 ngram の重みを 0 から 1 まで変化させた場合の翻訳品質（BLEU 指標）をグラフにしたものが図 5 である。BLEU は大きいほどよい翻訳性能を表しており、分岐 ngram を加えることにより、BLEU 指標を最大で 0.8 ポイント増加させることができることが分かった。

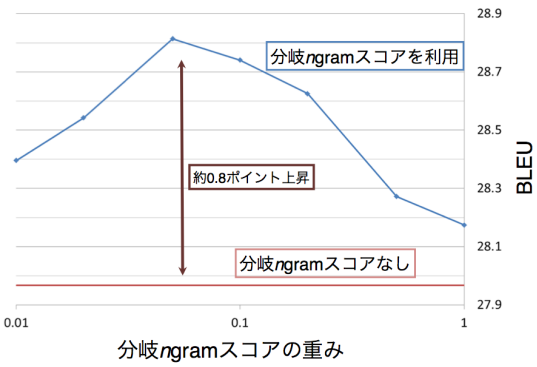


図 5 分岐 ngram による翻訳性能の改善

参考文献：

Lee,S. and Choi,K.S. (1998). “Automatic acquistion of language model based on headdependent relation between words.” In *Proc. of COLING-ACL 1998*, pp. 723-727.

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　１件）

① Chenchen Ding and Mikio Yamamoto. A generative dependency N-gram language model: unsupervised parameter estimation and application. 自然言語処理（査読有）. Vol.21, No.5, 2014. (印刷予定)

〔学会発表〕（計　１件）

① Chenchen Ding and Mikio Yamamoto. An unsupervised parameter estimation algorithm for a generative dependency N-gram language model. In Proc. of IJCNLP 2013（査読有）, pp.516-524, 2013.
http://lang.cs.tut.ac.jp/ijcnlp2013/

６．研究組織
(1)研究代表者
山本　幹雄（YAMAMOTO, Mikio）
筑波大学・システム情報系・教授
研究者番号：40210562

(2) 研究協力者
丁　塵辰（Ding, Chenchen）
筑波大学・大学院システム情報工学研究科
・博士後期課程

酒主　佳祐（SAKANUSHI, Keisuke）
筑波大学・大学院システム情報工学研究科
・博士前期課程

通事　寛奈（TOUJI, Hirona）
筑波大学・大学院システム情報工学研究科
・博士前期課程