(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3850742号
(P3850742)
(45)発行日　平成18年11月29日(2006.11.29)　(24)登録日　平成18年9月8日(2006.9.8)

(51)Int.Cl.		F I		
G10L 15/00	(2006.01)	G10L	15/00	200C
G06F 17/28	(2006.01)	G06F	17/28	U
G10L 15/18	(2006.01)	G10L	15/18	200D
G10L 15/06	(2006.01)	G10L	15/06	310Z

請求項の数 2　(全 10 頁)

(21)出願番号　特願2002-47047(P2002-47047)
(22)出願日　平成14年2月22日(2002.2.22)
(65)公開番号　特開2003-248496(P2003-248496A)
(43)公開日　平成15年9月5日(2003.9.5)
審査請求日　平成16年6月16日(2004.6.16)

(73)特許権者　393031586
株式会社国際電気通信基礎技術研究所
京都府相楽郡精華町光台二丁目２番地２
(74)代理人　100086391
弁理士　香山　秀幸
(72)発明者　中嶋　秀治
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内
(72)発明者　山本　博史
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内
(72)発明者　渡辺　太郎
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内

最終頁に続く

(54)【発明の名称】言語モデルの適応化方法

(57)【特許請求の範囲】
【請求項１】
音声翻訳器のタスクを拡大する際に、新規タスクに適応した第１の言語の言語モデルを作成するための言語モデルの適応化方法において、
第１の言語以外の第２の言語で記述された新規タスクの第１のモノリンガルコーパスを、機械翻訳器を用いて第１の言語に翻訳することによって、第１の言語で記述された新規タスクでの第２のモノリンガルコーパスを作成する第１ステップ、および
第１ステップで作成された新規タスクでの第２のモノリンガルコーパスに基づいて、言語モデルを適応化する第２ステップ、
を備えていることを特徴とする言語モデルの適応化方法。
【請求項２】
第２ステップは、第１ステップで作成された新規タスクでの第１の言語で記述された第２のモノリンガルコーパス、および一般タスクでの第１の言語で記述されたモノリンガルコーパスから、新規タスクに適応した第１の言語の言語モデルを作成するステップを備えていることを特徴とする請求項１に記載の言語モデルの適応化方法。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
この発明は、言語モデルの適応化方法に関する。
【０００２】

【従来の技術】
ある言語の言語モデルを適応先タスクに適応させるためには、小規模でも、適応先タスクでのコーパスであって、適応化しようとする言語モデルの言語で記述されたコーパスが必要となる。多言語の話し言葉の音声翻訳器の言語モデルのタスク適応となると、各言語の適応先タスクの小規模コーパスが必要となる。この収集はコスト高となり収集が困難である。
【０００３】
そこで、従来においては、集められた少量のコーパスを用いて言語モデルの適応化を行ったり（文献[1] 参照）、ＷＷＷ（World Wide Web) に情報を求めて得られた情報を用いて言語モデルの適応化を行ったりしていた（文献[2] 参照）。
【０００４】
文献[1] ：A. I. Rudnicky: "Language Modeling with Limited Domain Data," Proc. of the ARPA Spoken Language Systems Technogy Workshop,pp.66-69(1995).
文献[2] ：A. Berger, et al.:"Just-In-Time Language Modeling," Proc. of the ICASSP,pp. 705-708(1998).
【０００５】
これらは対象がディクテーションであり、話し言葉に比べて大量に存在する書き言葉を集めて利用する場合がほとんどであった。また、データの集めにくい医療所見のディクテーションや、マンマシンインターフェースの分野では、タスクの会話を記述するＣＦＧ(Context Free Grammar(文脈自由文法))を人手で作成し、作成したＣＦＧによって人工的に生成したデータを利用して言語モデルの適応を行っていた（文献[3] , [4] 参照）。
【０００６】
文献[3] ：伊藤伸泰, 荻野紫穂, 新島仁: " 文法を利用したＮ－ｇｒａｍモデルのタスク適応, ”言語処理学会第４回年次大会発表論文集, pp. 610-613(1998).
文献[4] ：Y.Wang,et al ed.,"A Unified Context-Free Grammar and N-gram Model for Spoken Language Processing," Proc. of ICASSP, 2000.
【０００７】
【発明が解決しようとする課題】
この発明は、適応化しようとする言語モデルの言語以外の言語で記述された新規タスクでのモノリンガルコーパスを用いることによって、新規タスクに適応した言語モデルを作成することができる言語モデルの適応化方法を提供することを目的とする。
【０００８】
【課題を解決するための手段】
請求項１に記載の発明は、音声翻訳器のタスクを拡大する際に、新規タスクに適応した第１の言語の言語モデルを作成するための言語モデルの適応化方法において、第１の言語以外の第２の言語で記述された新規タスクの第１のモノリンガルコーパスを、機械翻訳器を用いて第１の言語に翻訳することによって、第１の言語で記述された新規タスクでの第２のモノリンガルコーパスを作成する第１ステップ、および第１ステップで作成された新規タスクでの第２のモノリンガルコーパスに基づいて、言語モデルを適応化する第２ステップを備えていることを特徴とする。
【０００９】
請求項２に記載の発明は、請求項１に記載の言語モデルの適応化方法において、第２ステップは、第１ステップで作成された新規タスクでの第１の言語で記述された第２のモノリンガルコーパス、および一般タスクでの第１の言語で記述されたモノリンガルコーパスから、新規タスクに適応した第１の言語の言語モデルを作成するステップを備えていることを特徴とする。
【００１０】
【発明の実施の形態】
以下、図面を参照して、この発明の実施の形態について説明する。
【００１１】

〔１〕本発明の概要についての説明
ある言語の言語モデルを適応先タスクに適応させるためには、小規模でも、適応先タスクでのコーパスであって、適応化しようとする言語モデルの言語で記述されたコーパスが必要となる。多言語の話し言葉の音声翻訳器の言語モデルのタスク適応となると、各言語の適応先タスクの小規模コーパスが必要となる。この収集はコスト高となり収集が困難である。
【００１２】
本発明では、適応化しようとする言語モデルの言語以外の言語で記述された新たな適応先タスクでのモノリンガルコーパスに基づいて、言語モデルの適応化を行う。つまり、適応化しようとする言語モデルの言語以外の言語で記述された新たな適応先タスクでのモノリンガルコーパスを機械翻訳器によって翻訳することによって、適応化しようとする言語モデルの言語で記述された適応先タスクでの擬似的なコーパスを作成し、それを使って統計的言語モデルの適応化を行う。
【００１３】
機械翻訳器の翻訳用の知識には、隣接単語間の接続情報が保持されていることが期待できる。適応先タスクのモノリンガルコーパスを翻訳器によって翻訳することにより得られた疑似的なコーパスは、仮に訳文全体としては間違いを含んでいても、トピックや文のスタイルが適応先のタスクに適合し、かつ、隣接単語程度の局所的な文脈では適切な語順が反映されていると考えられる。したがって、作成された擬似的コーパスは言語モデルの適応用のコーパスとして利用できることが期待できる。
【００１４】
〔２〕言語モデルの適応化方法についての説明
はじめに、言語モデルの適応という課題を明らかにする。ここでは、ソース言語をLanguage2(例えば日本語）とし、ターゲット言語をLanguage1(例えば英語）とする。
【００１５】
図１に示すように、一般的な意味での言語モデルの適応という課題は、一般タスク(General Task)のデータと適応先タスク(Target Task) のデータとを使って、ターゲットのタスクに適応した言語モデルを作成することである。このため、小規模でも適応先のターゲットタスクのコーパスが必要となる。
【００１６】
一方、本発明での言語モデルの適応という課題は、図２に示すようになる。すなわち、図２でターゲット言語(Language1) の言語モデルの適応において、適応先タスク(Target Task) に対する言語データT_{L1}が存在しない場合に、その代わりとなる言語データT'_{L1}を、そのタスクでの他の言語のデータT_{L2}から言語翻訳によって作成し、それを使って言語モデルを適応させることである。
【００１７】
このターゲット言語(Language1) の言語データT'_{L1}の作成に用いられる翻訳器では、翻訳用の知識の中に、ターゲット言語の局所的な語順の情報が保持されている。そのため隣接単語間程度では比較的正しい語順の翻訳結果を得られる見込みがある。その結果、その作成された言語データT'_{L1}から、言語モデルに必要な隣接単語間の接続性に関する情報が得られる見込みがある。
【００１８】
適応の手法としては、ＭＡＰ適応(文献[5] 参照）やモデルの線形結合(上記文献[1] 参照）など様々な手法があるが、この実施の形態では、モデルの線形結合を利用する。
【００１９】
文献[5] ：H. Masataki,et.al: "Task Adaptation using MAP Estimation in N-gram Language Modeling," Proc. of the ICASSP, 1997,pp.783-786.
【００２０】
図３は、本発明による言語モデルの適応化方法の手順を示している。
ここでは、ターゲットタスク（新規タスク）に適応した第１の言語の言語モデルを生成す

る場合の言語モデルの適応化方法について説明する。
ステップ１：一般タスクのバイリンガルコーパスを使って機械翻訳器を作成する。ここで、一般タスクだけを使った一般タスク用の言語モデルLM_G（図２参照）も作成される。
【００２１】
ステップ２：ステップ１で作成した機械翻訳器を使って、第１言語以外の第２言語で記述されたターゲットタスクのモノリンガルコーパスT_{L2}（図２参照）を翻訳することにより、第１言語で記述されたターゲットタスクの擬似的なモノリンガルコーパスT'_{L1}（図２参照）を作成する。
【００２２】
ステップ３：ステップ２で作成された第１の言語で記述された擬似的なモノリンガルコーパスT'_{L1}を使って言語モデル適応用の言語モデル$LM_{T'L1}$（図２参照）を作成する。
【００２３】
ステップ４：ステップ１で作成した一般タスク用の言語モデルLM_Gと、ステップ３で作成した言語モデル適応用の言語モデル$LM_{T'L1}$とを結合することによって、ターゲットタスクに適応した第１言語の言語モデルを作成する。
【００２４】
なお、機械翻訳器としては、一般タスクのバイリンガルコーパスを使用して作成されたものが用いられているが、それ以外の方法で作成された機械翻訳器を用いてもよい。
【００２５】
また、言語モデル適応用の言語モデル$LM_{T'L1}$に結合される一般タスク用の言語モデルLM_Gとしては、ステップ１で機械翻訳器を作成する際に作成されたものが用いられているが、それ以外の方法で作成されたものを用いてもよい。
【００２６】
〔３〕機械翻訳器についての説明
【００２７】
上述したように、この発明による言語モデルの適応化方法においては、言語モデル適応用のデータを作成するために機械翻訳器が用いられる。さまざまな機械翻訳器の利用が可能であるが、後述する評価実験では翻訳の原理が明らかな統計的機械翻訳器を利用しているため、ここでは評価実験で用いられる統計的機械翻訳器の概要について説明する。
【００２８】
〔３－１〕統計的機械翻訳器の概要についての説明
評価実験で用いる翻訳器は、文献[6]のＩＢＭ　Ｍｏｄｅｌ４を基本とする統計的翻訳器である。
【００２９】
文献[6]：P. Brown, et. al,"The mathematics of statistical machine translation: Parameter estimation "Computational Linguistics, 19(2),pp.263-311(1993).
【００３０】
翻訳の問題は、雑音のある通信路での復号の問題と見做すことができる。例えば、日本語文（Ｊ）から英語文（Ｅ）への翻訳を考えると、日本語の文から英語の文への翻訳で最も尤もらしい翻訳結果（ここではこれをE^*とする）を得るという問題は、次式（１）として表現される。
【００３１】
$$E^* = \mathrm{argmax} P(E|J) \quad \cdots(1)$$
【００３２】
通常、この式をベイズの公式によって次式（２）に示すように変形し、ある１つの日本語の文を入力に定めた場合には分母が定数項となるので、分子のみで翻訳結果E^*の決定を行う。
【００３３】
$$E^* = \mathrm{argmax} P(J|E) P(E) / P(J) \quad \cdots(2)$$
【００３４】

ここで、分子のＰ（Ｊ｜Ｅ）は「翻訳モデル」、Ｐ（Ｅ）は「言語モデル」と呼ばれる。言語モデルとしては、隣接単語のＮ－ｇｒａｍがしばしば用いられる。

【００３５】

これらのモデルのパラメータは全て学習用の一般タスクのバイリンガルコーパスから統計的に推定される。これらのモデルを用いることにより、訳語を選択し、Ｐ（Ｅ）の制約のもとで語順を整えて、翻訳を達成する。そのため、対象が新たなタスクの場合でも、モデルのパラメータが正しく推定されていれば、出力される翻訳結果の中に、局所的な文脈では比較的正しい単語並びが得られ、言語モデル適応用のデータとして利用できることが期待できる。

【００３６】

本発明では、機械翻訳器で生成されたテキストから作られた言語モデル適応用の言語モデルと、一般タスクのコーパスから作られた言語モデル（この実施の形態ではＰ（Ｅ））との線形結合により、タスクに適応化された言語モデルを作成する。

【００３７】

〔４〕評価実験

【００３８】

以下、本発明の有効性を翻訳先言語のターゲットタスクでのテストセットパープレキシティの削減によって確認する。これに用いるデータ、実験条件について説明する。

【００３９】

〔４－１〕一般およびターゲットタスクのデータ

本実験では、日英の対訳コーパス（以後、「フレーズブックコーパス」と呼ぶ）を利用する。文の内容は旅行時の会話表現である。

【００４０】

これらの表現は、あらかじめ人手によって、「空港」、「飛行機内」、「レストラン」などの場面を主とした複数のカテゴリに分類されている。分類カテゴリの例を表１に示した。

【００４１】

【表１】

基本	空港	飛行機
両替	宿泊	レストラン
軽食	飲食	移動
買物	観光	美容
連絡	帰国	コミュニケーション
トラブル	ホームステイ	留学
ビジネス	研究	

【００４２】

カテゴリーの中の「空港」での会話表現を評価実験のターゲットタスクとして設定し、残りのカテゴリーを一般タスクとして設定した。これらの内訳は、表２の通りである。

【００４３】

【表２】

コーパス名	文数	単語数
General	152,857	1,197,691
t1000	1,000	7,269
t2000	2,000	15,415
t4739	4,739	36,737
test	4,739	36,191

【００４４】
ターゲットタスクのコーパスの規模と適応の効果との関係を調べるために、サイズの異なるターゲットタスクコーパスを数通り用意した。表２の”General ”を一般タスクとし、”t1000 ”, ”t2000 ”, ”t4739 ”を量が異なる３つのターゲットタスクのコーパスとし、”test”をターゲットタスクの評価用コーパスとする。なお、文と単語の総数は英語で数えた結果である。
【００４５】
機械翻訳器へは、”t1000 ”, ”t2000 ”および”t4739 ”の形態素解析済みの日本語形態素列を入力する。そして、出力された翻訳結果の英語文（単語への分割済）を適応用のデータT'_{L1}（図２参照）として利用する。
【００４６】
〔４－２〕実験の手順
ここでは、翻訳器によって作成されたテキストを用いて言語モデルの適応化を行い、そのモデルの性能をパープレキシティーの削除量で評価する。
【００４７】
最初に、翻訳器によって完全な訳文（本実験では英文）が作成される場合を想定して、ターゲットタスクの３種類のコーパスのそれぞれに対応するフレーズブックコーパス内の訳文（ここでは英語文）を使って言語モデルの適応化を行った場合のパープレキシティーの削減量を調査する。
【００４８】
次に、翻訳器によって翻訳されたデータを言語モデルの適応に用いたことの効果（本発明の効果）を調査する。
【００４９】
まず、上記実施の形態で説明した手順にしたがって、統計的機械翻訳器の翻訳モデルと言語モデルとを作成する。つまり、一般タスクのバイリンガルコーパス（”General ”の両言語）のみを使用して、翻訳器とターゲット言語（本実験では英語）の言語モデルLM_G（図２参照）とを作成する。
【００５０】
次に、ターゲットタスクのモノリンガルコーパス（本実験では、日本語のt1000 , t2000 , t4739 ）のみを上記機械翻訳器で翻訳し、得られた結果を使用して、ターゲット言語(本実験では、英語）でのターゲットタスク用の言語モデルLM_T （添字のTは、t1000 , t2000 , t4739 を表す）を作成する。
【００５１】
最後に、２つの言語モデルLM_G とLM_T とを線形結合することにより、適応化した言語モデルを作成する。本実験では、言語モデルには単語３gramを、線形結合の結合係数は削除補間法（上記文献[1] 参照）によって決めた。
【００５２】

本実験は、言語モデルの新たな適応先となるターゲットタスクにだけ出現する単語とその訳語の対が、機械翻訳器に存在しない設定の元で行った。

【００５３】

〔４－３〕実験結果と考察

【００５４】

以下に示す結果は全て、翻訳先言語のターゲットタスクのテストセット（本実験では「空港」タスクの英語）での単語パープレキシティー（ＰＰ値）またはその削減率で示す。

【００５５】

〔４－３－１〕対訳コーパスの理想訳を使った場合（性能の上限）

ターゲットタスクのデータを仮に集めることができた場合に、どこまでＰＰ値の削減が得られるかという性能の上限を確認しておくために、対訳コーパスの訳を使って適応した場合のＰＰ値を計算した。

【００５６】

この結果を、表３の２～４行目（”General ＋t ＜サイズ＞”）に示す。同じ表３の１行目（”General only”）は、一般タスクのバイリンガルコーパスだけから作られた言語モデルでのＰＰ値である。

【００５７】

【表３】

コーパスの組み合わせ	ＰＰ	削減率[%]
General only	32.0	base
General + t1000	23.5	26.7
General + t2000	21.8	31.9
General + t4739	19.8	38.1

【００５８】

このように、一般タスクの言語モデルとターゲットタスクの言語モデルとの適応によって作成された言語モデルにおけるＰＰ値のほうが、他に比べて小さくなっていることがわかる。

【００５９】

また、図４の実線Ａは、ターゲットタスクのコーパスを含まない一般タスクのバイリンガルコーパスの増加に伴うＰＰの変化を示すグラフであり、破線Ｂは一般タスクのバイリンガルコーパスにターゲットタスクのバイリンガルコーパス内の訳文（英文）を順次追加した場合のＰＰの変化を示すグラフ（表３のGeneral ＋t ＜サイズ＞でのＰＰ値）である。図の横軸は文数（コーパスサイズ）を、縦軸はＰＰ値を示している。

【００６０】

図４から、一般タスクのコーパスをさらに倍の規模だけ集める（実線の右側への延長）よりもターゲットタスクのデータを適応させたほう（破線）がＰＰ値の削減の効果が大きくなっていることがわかる。

【００６１】

以上２点から、本データでは、言語モデル適応の方が有効であることが観察された。

【００６２】

これらの結果から、本実験の設定では、翻訳が１００％成功すれば、適応先のデータが増加するに伴って、適応先のタスクでのオープンなテストセットのＰＰ値が削減し、最大３８．０％近くまで相対的にＰＰ値を下げられる可能性があることがわかる。

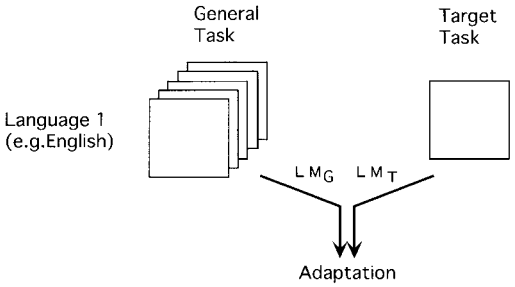
【００６３】

〔４－３－２〕翻訳されたテキストを用いる場合
次に、理想的な翻訳結果を用いる代わりに、実際に機械翻訳器を使って作成されたターゲットタスクのデータを適応に使った場合のＰＰ値を、各組合わせについて計算した結果を表４に示す。
【００６４】
【表４】

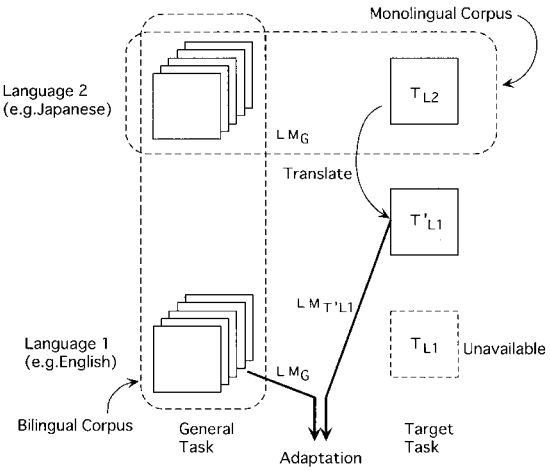
コーパスの組み合わせ	ＰＰ	削減率[%]
General only	32.0	base
General + t1000	27.8	13.1
General + t2000	27.9	12.6
General + t4739	27.9	12.8

【００６５】
表４から、機械翻訳器によって作成された、新しいターゲットタスクに対するテキストを言語モデルの適応に用いることにより、適応前のＰＰ値からの１３％のＰＰ値の削減が得られることがわかる。
なお、図５の実線Ｂは、図４の破線に相当するものであり、一般タスクのバイリンガルコーパスにターゲットタスクのバイリンガルコーパス内の訳文（英文）を順次追加した場合のＰＰの変化を示すグラフであり、破線Ｃは、一般タスクのバイリンガルコーパスにターゲットタスクのバイリンガルコーパスの日本文を機械翻訳した文（英文）を順次追加した場合のＰＰの変化を示すグラフである。
【００６６】
【発明の効果】
この発明によれば、適応化しようとする言語モデルの言語以外の言語で記述された新規タスクでのモノリンガルコーパスを用いることによって、新規タスクに適応した言語モデルを作成することができるようになる。この結果、各言語の小規模コーパスを集めなくて済み、そのためコストがかからなくなる。
【図面の簡単な説明】
【図１】一般的な意味での言語モデルの適応という課題を説明するための模式図である。
【図２】本発明での言語モデルの適応という課題を説明するための模式図である。
【図３】本発明による言語モデルの適応化方法の手順を示すフローチャートである。
【図４】ターゲットタスクのコーパスを含まない一般タスクのバイリンガルコーパスの増加に伴うＰＰの変化と、一般タスクのバイリンガルコーパスにターゲットタスクのバイリンガルコーパス内の訳文（英文）を順次追加した場合のＰＰの変化とをそれぞれ示すグラフである。
【図５】一般タスクのバイリンガルコーパスにターゲットタスクのバイリンガルコーパス内の訳文（英文）を順次追加した場合のＰＰの変化と、一般タスクのバイリンガルコーパスにターゲットタスクのバイリンガルコーパスの日本文を機械翻訳した文（英文）を順次追加した場合のＰＰの変化とをそれぞれ示すグラフである。

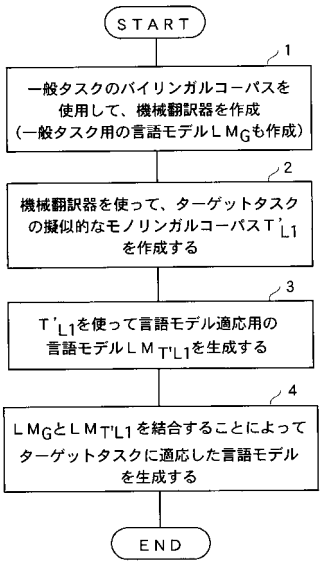
【図１】


General
Task
Target
Task
Language 1
(e.g.English)
LMG
LMT
Adaptation

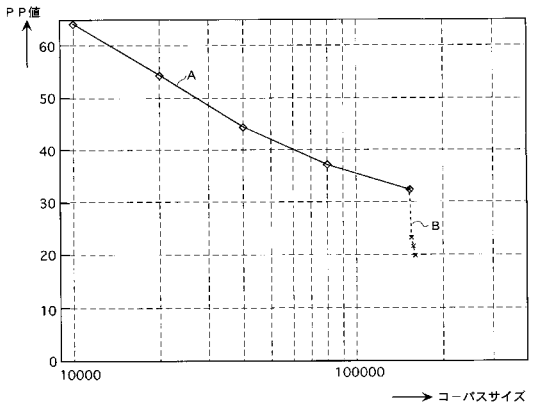
【図２】


Monolingual Corpus
Language 2
(e.g.Japanese)
TL2
LMG
Translate
T'L1
LMT'L1
Language 1
(e.g.English)
TL1
Unavailable
LMG
Bilingual Corpus
General
Task
Adaptation
Target
Task

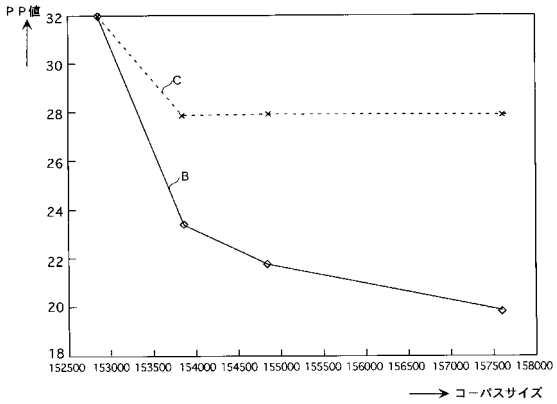
【図３】


START
1
一般タスクのバイリンガルコーパスを
使用して、機械翻訳器を作成
(一般タスク用の言語モデルLMGも作成)
2
機械翻訳器を使って、ターゲットタスク
の擬似的なモノリンガルコーパスT'L1
を作成する
3
T'L1を使って言語モデル適応用の
言語モデルLMT'L1を生成する
4
LMGとLMT'L1を結合することによって
ターゲットタスクに適応した言語モデル
を生成する
END

【図４】


PP値
60
50
40
30
20
10
0
A
B
10000
100000
コーパスサイズ

【図５】


PP値
32
30
28
26
24
22
20
18
C
B
152500 153000 153500 154000 154500 155000 155500 156000 156500 157000 157500 158000
コーパスサイズ

フロントページの続き

審査官　樫本　剛

(56)参考文献　特許第２６４４１７１（ＪＰ，Ｂ２）
特許第３３０５９５３（ＪＰ，Ｂ２）
特開２０００－３０５９３０（ＪＰ，Ａ）
特開２００１－３４３９９３（ＪＰ，Ａ）
特開２０００－２１４８８１（ＪＰ，Ａ）
政瀧他，最大事後確率推定によるN-gram言語モデルのタスク適応，電子情報通信学会論文誌　D-II，日本，１９９８年１１月２５日，Vol.J81-D-II、No.11，p.2519-2525
P. E. Brown et al., The Mathematics of Statistical Machine Translation : Parameter Estimation, Computational Linguistics, 米国, １９９３年　６月, Vol.19, No.2, p.263-311

(58)調査した分野(Int.Cl.，ＤＢ名)
G10L 15/00
G10L 15/06
G10L 15/18
G06F 17/28