文クラスタリングを用いた言語モデル学習データの選択手法

安田 圭志[†,‡]　山本博史[†]　隅田 英一郎[†,‡]

†情報通信研究機構音声言語グループ

‡ATR 音声言語コミュニケーション研究所

〒619-0288「けいはんな学研都市」光台二丁目 2 番地 2

E-mail:　†{keiji.yasuda, hirofumi.yamamoto.eiichiro.sumita}@nict.go.jp

あらまし　言語資源の整備，開発の進展に伴い，現在非常に大規模な言語コーパスが利用可能となっている．このような大規模なコーパスから統計的言語モデルを学習する際，コーパスの大規模化により得られる言語モデルの性能が向上するというメリットがある反面，言語モデル学習に要する処理時間が非常に長くなるという問題が生じる．本研究ではこのような言語モデル学習時の計算負荷の問題を解決するため，大規模な言語コーパスの中から対象とするタスクと異なるデータや，雑音的なデータを除去することにより，得られる言語モデルの性能を担保しつつ学習データの量を減らし，処理時間の問題を軽減させる手法を提案する．提案手法に関する実験の結果では，学習セットのサイズを 60%程度削減しつつ，言語モデルの性能をも改善することが可能となった．

1．はじめに

情報処理技術の進展や，インターネットの普及に伴い，多くの文書や情報が電子化されるようになった．このようなデータを整備・蓄積することにより，現在では非常に大規模な言語コーパスが開発され，利用することが可能となってきている．

このような大規模な言語コーパスは，自然言語処理の分野においても様々な研究に利用可能であるが，本研究では，統計的翻訳，形態素解析，音声認識等においても幅広く応用される統計的言語モデルについて扱う．

ここで，統計的言語モデルと利用可能な言語コーパスの規模との関係について見ると，コーパスの大規模化により，言語モデルの性能が向上するというメリットがある反面，言語モデルの学習に要する処理量や，処理の際に必要となるメモリ量が増大するという問題が生じる．

本研究では，このような問題を解決するため，大規模な言語コーパスの中から，対象とするタスクと異なるデータや，雑音的なデータを除去することにより，得られる言語モデルの性能を担保しつつ，学習データの量を減らし，言語モデルの学習に要する計算機的負担を軽減させる手法を提案する．

以下では，2 節提案手法について説明し，3 節で実験結果について述べる．最後に 4 節で論文を結ぶ．

2．提案手法

提案手法では，小規模な開発用セットと，大規模な言語コーパスとを必要とする．開発セットは，言語モデルを用いるアプリケーションにおいて対象とするタスクに属する文からなる小規模コーパスである．一方，大規模言語コーパスは，ある特定のタスクに属する文だけではなく，種々のタスク，ドメインに属する文からなるコーパスである．

提案手法の考え方は，文クラスタリングにより大規模コーパスを特性の近いものごとのサブセット(クラスタ)に分け，次に，各サブセットと開発セットの類似性をパープレキシティーにより測定し，開発セット

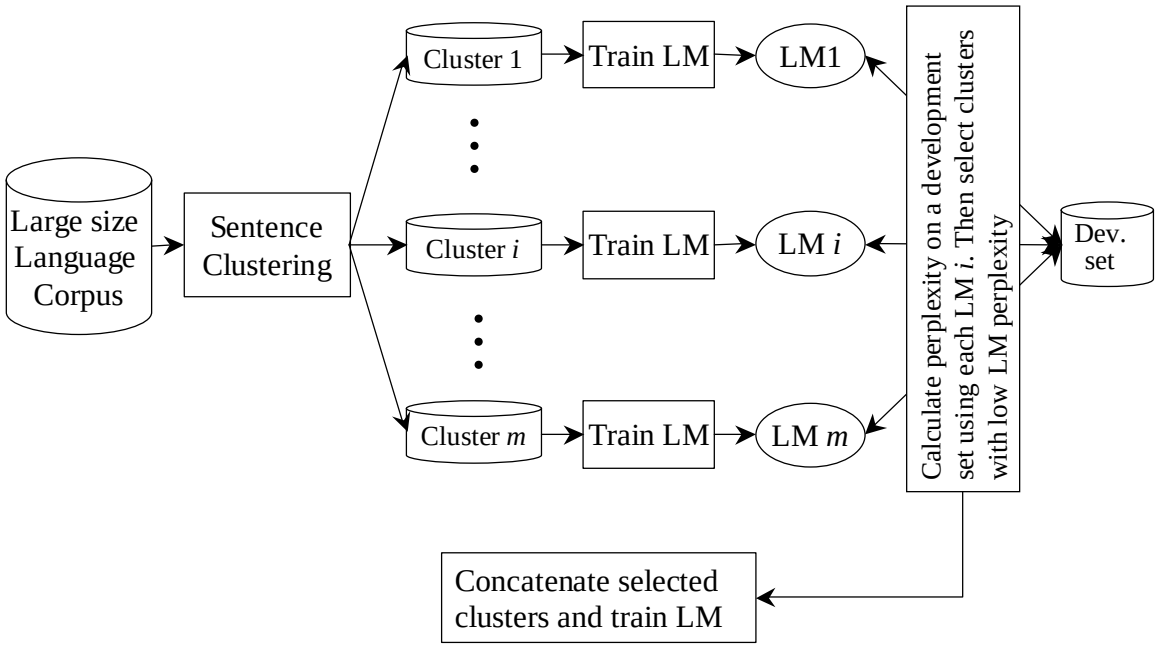

図 1　提案手法の処理の流れ

表 1　実験に用いたデータの詳細

Data type	Size (Word)	Explanation
Large size corpus	382m	English corpus from LDC (LDC2002E18, LDC2003E07, LDC2003T17, LDC2004T07, LDC2005T06, LDC2002T01, LDC2003E14, LDC2004E12, LDC2004T08, LDC2005T10 and part of LDC2005T12)
Development set	17k	English Translation of Chinese version of "Voice of America" broadcast news (Two translations per one sentence)
Test set	61k	English Translation of Chinese version of "Voice of America" broadcast news (Two translations per one sentence)

に類似したサブセットを言語モデルの学習セットとして用いるというものである．

図 1 に提案手法の処理の流れを示す．提案手法では，以下の手順により学習データの選択を行う．

(1)対象とする大規模言語コーパスに対して，文クラスタリング[1]を適用し，大規模コーパスを m 個のサブセットに分割する．

(2)(1)で得られた全てのサブセットに対して言語モデルを学習する．

(3)(2)で得られたそれぞれの言語モデルを用いて，開発セットのパープレキシティーを計算する．

(4)(3)で得られたパープレキシティーの値が閾値よりも小さくなるサブセットのみを集め，これらを学習セットとする．

3．実験結果

3.1 実験条件

実験において，開発セットと評価に用いるテストセットは， TC-STAR (Technology and Corpora for Speech to Speech

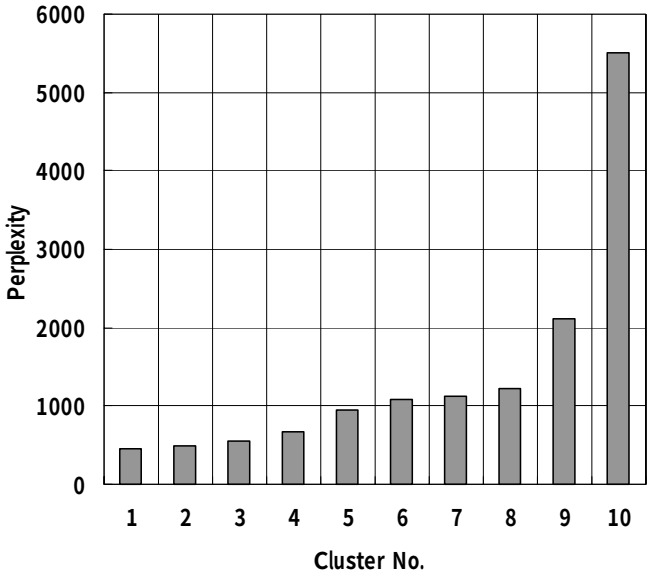

図 2　クラスタ毎に学習した言語モデルにおけるパープレキシティー

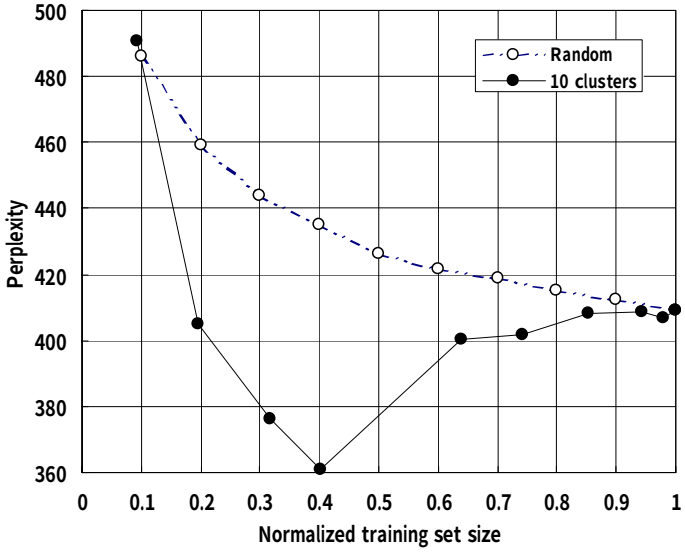

図 3　提案手法により得られたテストセットパープレキシティー

Translation)[2]第二回評価キャンペーンでのデータを用い，大規模言語コーパスとしては，LDC コーパスを用いた．これらのデータの詳細を表 1 に示す．

本実験において，文クラスタリング時は 1-gram の言語モデルを，文クラスタリング以外では，バックオフスムージング[4]とグッド・チューリング法[5]を用いた 3-gram の言語モデルを用いた．各言語モデルの性能評価は，表 1 で示したテストセットに対するパープレキシティーにより行なう．

3.2 実験結果

図 2 は，文クラスタリングの結果をもとに，各クラスタ毎に学習した言語モデルの開発セットに対するパープレキシティーである．文クラスタリングの際のクラスタ数である m の値は 10 としている．図 2 において，縦軸はパープレキシティーを表し，横軸の番号はクラスタ番号を表している．このクラスタ番号はパープレキシティーの値が小さい順に 1～10 としている．

次に，図 2 で示した結果を用い，学習セットの選択を行なった結果について述べる．図 3 は，図 2 で示したクラスタ 1～10 を順々に結合して学習セットとし，テストセットパープレキシティーの変化を調べた結果である．図 3 において縦軸はテストセットパープレキシティーを表し，横軸は選択された学習セットのサイズ表す．横軸の値は，選択前の大規模コーパスのサイズが 1 となるように単語数で正規化した値である．また，図 3 において，●は提案手法により学習セットの選択を行った結果を表し，○は比較のため，ランダムに大規模言語コーパスから学習セットを選択した結果[1]を表す．

図 3 を見ると，ランダムに選択した場合は，学習セットのサイズが大きくなるにつれて，テストセットパープレキシティーの値も小さくなっていることが分かる．一方，提案手法により学習セットを選択すると，学習セットのサイズが全体の 4 割程度の場合（クラスタ番号 1～4 を利用した場合）に最もテストセットパープレキシティーが小さくなり，以降，学習セットのサイズが大きくなるにつれ，テストセットパープレキシティーの値も大きくなっている．前述の提案手法により最もパープレキシティーが低くなる点と，大規模言語コーパスを全

[1] 各プロットは，学習セットのランダム選択 3 試行の加算平均を取っている．

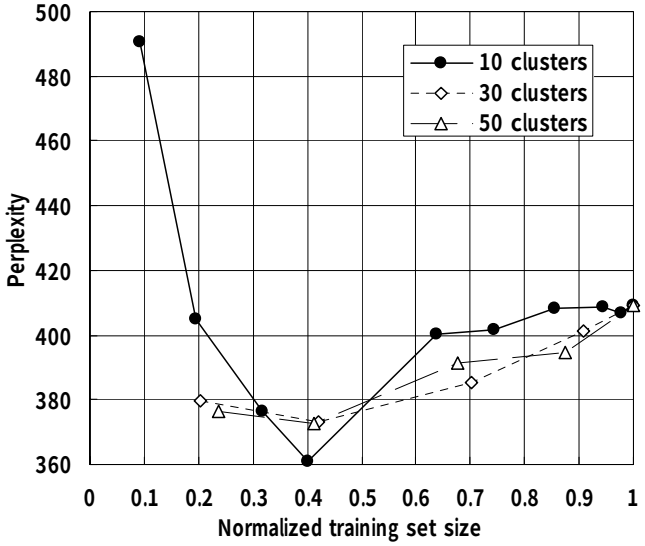

図 4　提案手法により得られたテストセットパープレキシティー
(クラスタ数が 10, 30, 50 の場合)

て用いた場合とを比較すると，提案手法により得られるテストセットパープレキシティーの値の方が約 12%小さくなっており，60%の学習セットのサイズ削減が可能となっただけでなく，言語モデルの性能も改善されていることが分かる．

3.3 考察

クラスタリングにおいて，適切なクラスタ数(m)を調べるため，m の値を 10，30，50 とした場合の実験結果を図 4 に示す．図 4 において，縦軸と横軸は図 3 と同様に，テストセットパープレキシティーと学習データのサイズを表している．図 3 を見ると，全ての場合において学習セットを全体の 4 割程度にした場合にパープレキシティーの値が最も小さくなっている．学習セットを 4 割程度にした場合について，各々の m の値での結果を比較すると，m が 10 の場合にパープレキシティーの値が最も小さくなっていることが分かる．

図 4 より今回の実験の範囲内では，m の値が 50 の場合に最も良い選択が可能となることが分かった．

4. まとめ

大規模言語コーパスから，言語モデル学習に用いる学習セットを選択する方法を提案した．LDC コーパスと TC-STAR 第二回評価キャンペーンのデータを用いた実験の結果，提案手法により学習セットのサイズを 60%程度削減することが可能となり，言語モデル学習に必要となる処理時間を短縮するだけでなく，言語モデルの性能を改善することも可能となった．

今後の検討課題として，本研究で得られたモデルの小規模化と改善が，音声認識や機械翻訳などのアプリケーションレベルにおいて，どの程度の影響があるかを調べる必要がある．

文　献

[1] D. Carter, “Improving Language Model by Clustering Training Sentences," Proc. ACL, pp.59-64, 1994.

[2] http://www.elda.org/en/proj/tcstar-wp4/tcs-run2.htm

[3] http://www.ldc.upenn.edu/

[4] S. M. Katz, “Estimation of probabilities from sparse data for the language model component of a speech recognizer,” IEEE Transactions on Acoustics, Speech and Signal Processing, pp.ASSP-35(3), 400-401, 1987.

[5] I. J. Good, “The population frequencies of species and the estimation of population parameters,” Biometrika, pp.40(3), 237-264, 1953.