Vol.2011-SLP-85 No.14
2011/2/5

WWW から得られる Term Frequency 情報に基づく PLSA 言語モデル

宮崎将隆[†]　　川端豪[†]

PLSA は学習データ内における単語の出現頻度を基にトピックをモデル化する手法である．しかし，話題判定を tf だけで行うより，tfidf で判定した方が効果的であり，パープレキシティ削減も期待できる．そこで，本報告では WWW から得られる tfidf 統計に基づいた PLSA 言語モデルについて検討する．このシステムでは WWW から 1 万語の語彙に対して，idf 値を計算している．実験の結果，本手法では通常の PLSA 言語モデルよりもテストセットパープレキシティを削減することができ，トピック判定が改善されたと考える．

Improvement of PLSA Language Models for Perplexity Reduction based on the World Wide tfidf Statistics.

Masataka Miyazaki[†]　and Takeshi Kawabata[†]

PLSA (Probabilistic Latent Semantic Analysis) is a method for topic modeling based on the Term Frequency of the word in a training text. However, the tf values in limited data may not be reliable enough to reduce perplexities for speech recognition. This paper describes an improvement of PLSA language model based on the world wide “tfidf” statistics. The system searches for 10,000 vocabulary words in World Wide Web, and calculates general idf values for them. Experiments show that the proposed method effectively achieves accurate topic identification, and reduces the task-set perplexity.

1. はじめに

大語彙連続音声認識の言語モデルとして，一般的には N-gram モデル[1]が用いられる．N-gram モデルでは，直前の（N-1）個の単語によって次に来る単語を予測するという仕組みになっている．しかし，人が話をする際にはある「トピック」にのっとり，1 文もしくは 1 講演といったもっと長い時間を有するため，直前の数単語だけではなく，もっと長距離の文脈情報を次単語の予測モデルに組み込むことで性能を向上させることができる．

言語モデルを目的のトピックに適応する手法としては，認識対象に類似したコーパスを収集し言語モデルを構築する手法や複数のトピックが混在するコーパスからトピックの推定を行う手法がある．前者の研究では，梶浦ら[2]や増村ら[3]のように有効な検索クエリを構築し，WWW からコーパスを収集する手法がある．一方後者では，確率文法とマルコフモデルによる話題制御の方法[4]や政瀧らが提案した MAP 推定を用いた N-gram 言語モデルのタスク適応[5]などの研究がなされている．

また，トピックをモデル化する手法として，Thomas Hohmann によって提案された PLSA 言語モデル[6]がある．PLSA 言語モデルは話題を判定し，単語の出現確率に重み付けを行う手法である．学習最適化の方法[7]や，語彙分割[8]の発展研究がなされている．しかし，PLSA 言語モデルでは話題判定に用いる情報として学習データから得られる単語の出現頻度のみを用いている．情報検索の分野においては文書間に渡る情報である文書頻度を組み合わせた指標である tfidf [9]が高いキーワード選択の指標として知られている．

そこで本報告では，PLSA のトピック推定のための学習に用いる単語の出現頻度を学習データから計算した単語出現頻度だけでなく，WWW から計算した文書頻度情報を加えた情報によって代替するという方法を提案する．

そして，本手法の有効性を通常の PLSA 言語モデルとのパープレキシティの比較により検証する．

2. PLSA 言語モデル

今回提案する手法のベースとなる PLSA 言語モデルについて簡単に説明しておく．

PLSA（Probabilistic Latent Semantic Analysis）とは，学習データ内における単語の出現頻度を基に，トピックをモデル化する手法である[6][8].

[†] 関西学院大学　理工学研究科
School of Science and Technology, Kwansei Gakuin University

Vol.2011-SLP-85 No.14
2011/2/5

文脈情報 h を反映した単語 w の出現確率は式(1)で与えられる．

$$P(w \mid h) = \sum_{z \in Z} P(z \mid h) P(w \mid z) \qquad (1)$$

$P(z \mid h)$ は混合比として見ることができ，$P(w \mid z)$ は内部 unigram モデル（あるトピックでの各単語の出現確率）が与える確率である．そのため，PLSA 言語モデルの実体は，目的のトピックに則して複数の内部 unigram を適した混合比で組み合わせた unigram の複数混合モデルであると言える．

この混合比と内部 unigram を学習データにより，式(2)を最大化するように反復学習（Tempered EM アルゴリズム）を行う．

$$l(\theta; N) = \sum_{w \in W} \sum_{d \in D} n(d, w) \log \sum_{z \in Z} P\{z \mid d) P(w \mid z) \qquad (2)$$

$n(d, w)$ ：データ d 中の単語 w の出現回数

パラメータ推定のための反復学習には，式(3)～(6)の式を用いて行う．

E-step：

$$P^{(k)}(z \mid d, w) = \frac{\left\{P^{(k)}(z) P^{(k)}(d \mid z) P^{(k)}(z \mid d, w)\right\}^{\beta}}{\sum_{z \in Z} \left\{P^{(k)}(z) P^{(k)}(d \mid z) P^{(k)}(z \mid d, w)\right\}^{\beta}} \qquad (3)$$

M-step：

$$P^{(k+1)}(w \mid z) = \frac{\sum_{d \in D} n(d, w) P^{(k)}(z \mid d, w)}{\sum_{w \in W} \left\{\sum_{d \in D} n(d, w) P^{(k)}(z \mid d, w)\right\}} \qquad (4)$$

$$P^{(k+1)}(d \mid z) = \frac{\sum_{w \in W} n(d, w) P^{(k)}(z \mid d, w)}{\sum_{d \in D} \left\{\sum_{w \in W} n(d, w) P^{(k)}(z \mid d, w)\right\}} \qquad (5)$$

$$P^{(k+1)}(z) = \frac{\sum_{w \in W} \sum_{d \in D} n(d, w) P^{(k)}(z \mid d, w)}{\sum_{w \in W} \sum_{d \in D} n(d, w)} \qquad (6)$$

この E-step と M-step を交互に繰り返すことで，式(2)を最大化するようなモデルを生成することができる．

目的のトピックへの適応には、$P(w \mid z)$ は文脈情報に依存しないので学習で得られた値を固定し，適応したいテキストの単語出現頻度 $n(h, w)$ に対して、学習時と同じように Tempered EM アルゴリズムによって、PLSA 言語モデルが尤度最大化する混合比を推定することで行う．その際に必要な式は下式のようになる．

E-step：

$$P^{(k)}(z \mid h, w) = \frac{\{P^{(k)}(z) P^{(k)}(h \mid z) P^{(k)}(w \mid z)\}^{\beta}}{\sum_{z \in Z} \{P^{(k)}(z) P^{(k)}(h \mid z) P^{(k)}(w \mid z)\}^{\beta}} \qquad (7)$$

M-step：

$$P^{(k+1)}(h \mid z) = \frac{\sum_{w \in W} n(h, w) P^{(k)}(z \mid h, w)}{\sum_{z \in Z} \{\sum_{w \in W} n(h, w) P^{(k)}(z \mid h, w)\}} \qquad (8)$$

3. tfidf

今回提案する手法では tfidf [9]の概念を使用しているので, tfidf についてここで説明しおく．

tfidf は情報検索の分野で開発された話題判定に対するキーワード選択の指標の 1 つである．計算式は式(9)で与えられる．

$$tfidf = tf(w,d) \cdot idf(w) = tf(w,d) \cdot \log \frac{N}{df(w)} \qquad (9)$$

$tf(w,d)$ ：文書 d における単語 w の出現頻度

N ：文書集合の数

$df(w)$ ：文書集合中での単語 w の出現した文書数

このような計算式で求められる指標であるため, 出現回数の多い単語の $tf(w,d)$ は大きくなり $tfidf$ も大きくなる．また，どの文書にも出現する単語の $df(w)$ は大きくなり $idf(w)$ は小さくなるため，$tfidf$ も小さくなる．

このように，単語の出現頻度によって，何度も繰り返し言及される単語の重要度を上げることと，文書頻度によって，どの文書にも出現するような一般的な単語（助詞等）の重要度を下げることを併用しており，高いキーワード選択の指標として知られている．

4. WWW から得られる tfidf 情報に基づく PLSA 言語モデル

本報告で提案する tfidf の概念を組み込んだ PLSA 言語モデルについて説明する．

4.1 パープレキシティを削減するために PLSA を使う時の問題点

通常の PLSA 言語モデルでは，言語モデルを構築する際の話題判定に各学習データから得られる単語出現頻度を用いている．

図 1 は各学習データとベクトル空間を表している．w_1 ，w_2 ，w_3 は語彙単語の種類である．単語の出現頻度をベクトルの要素として考えると，図 1 のように各データは全語彙数を次元とするベクトル空間上において一点（○印）を指し，単語の出現頻度が類似したデータ同士は距離が近くなる．そのため，ベクトル間の距離が近いものは

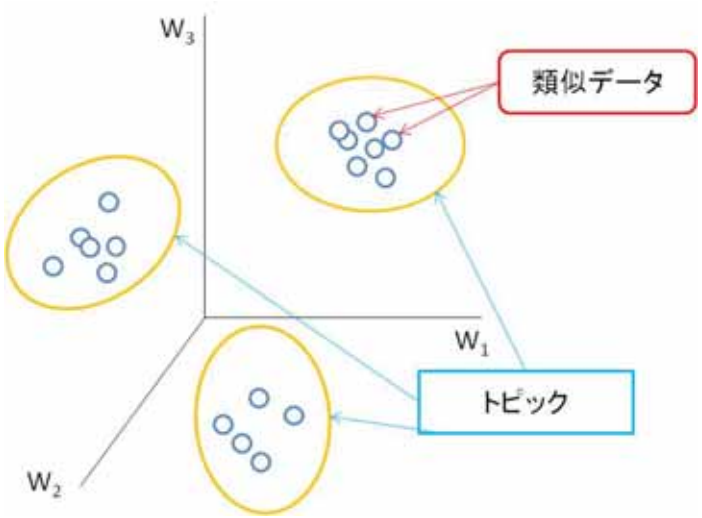

図 1：各学習データとベクトル空間

類似トピックと判定されていることになる．

ここで，全語彙を次元としたベクトル空間を考えるということはすべての単語が話題の判定において同等であるとみなしていることになる．しかし，話題判定において有効でないと考えられる，話し言葉に頻出する「えー」や「です」といった一般的な単語や「て」「に」「を」「は」などの助詞は出現回数が自然と多くなり，ベクトル空間において各データの位置を特徴づけてしまう．そのため，本来トピックの違うデータ同士であるのに 2 つが類似していると誤判定してしまう. 一般的に, 2 つ以上の unigram を混合する際には，元の unigram が持つ分布を歪めてしまうため，パープレキシティの劣化が起こる．その影響を最小限に抑えるために，分布の類似したもの（トピックが同じもの）同士を混合していく必要がある．つまり，話題判定の誤りはそのままパープレキシティの劣化につながるのである．

そこで，単語自身の持つ話題判定への貢献性を考慮に入れることを考える．すべての単語を同等に扱い各データの類似度を比較するよりも，tfidf に基づいて補正を加えた後に比較を行うことで，話題判定の精度は向上すると考えられる．

このように，類似度の計算を行う際に，すべての単語を同等に考えてしまうと，話題判定に影響を及ぼさない単語による誤判定が伴うので，tfidf によって単語間に話題判定に対する重要度の差をつけることで，この問題を解消していくことを考える．

4.2 提案方法

本報告では前節に述べた問題点を改善するために出現回数 $n(d,w)$ に，その単語のトピック依存性である idf 値を加味することを考えた．この idf 値は WWW 上で単語の使用統計を行うことで算出する．情報検索の分野において，キーワード選択の指標には単語の出現頻度のみを扱うだけではなく，文書頻度も組み合わせた指標である tfidf を用いる方法が提案されており，PLSA 言語モデルにおける話題判定に対しても性能の向上が期待できる．

具体的には，検索エンジンにインデックスされているすべての Web ページを母集団とみなし，ある単語を検索クエリとし検索 API にリクエストを行うことで得られる WWW のヒット件数を文書頻度とみなすことで，ある単語の WWW における idf を計算し，この値を PLSA 言語モデルの学習に用いる式(4)(5)(6)の $n(d,w)$ に乗算することで内部 unigram の推定に tfidf の概念を組み込むというものである．その際の式は(13)～(15)になる．

$$P^{(k+1)}(w \mid z) = \frac{\sum_{d \in D} n(d,w) \cdot idf(w) P^{(k)}(z \mid d,w)}{\sum_{w \in W} \left\{ \sum_{d \in D} n(d,w) \cdot idf(w) P^{(k)}(z \mid d,w) \right\}} \quad (13)$$

$$P^{(k+1)}(d \mid z) = \frac{\sum_{w \in W} n(d,w) \cdot idf(w) P^{(k)}(z \mid d,w)}{\sum_{d \in D} \left\{ \sum_{w \in W} n(d,w) \cdot idf(w) P^{(k)}(z \mid d,w) \right\}} \quad (14)$$

$$P^{(k+1)}(z) = \frac{\sum_{w \in W} \sum_{d \in D} n(d,w) \cdot idf(w) P^{(k)}(z \mid d,w)}{\sum_{w \in W} \sum_{d \in D} n(d,w) \cdot idf(w)} \quad (15)$$

5. 検証実験

通常の PLSA 言語モデルと提案手法のテストセットパープレキシティを比較する．

5.1 実験条件

データには CSJ（日本語話し言葉コーパス）を用いた．全データ数は 3302 講演であり，この内の 15 講演をランダムに抜き出し評価用データとし，残りの 3287 講演を言語モデルの学習に用いた．

PLSA 言語モデルを構築する際の語彙数は 1 万とし，内部 unigram 数は 50 とした．また，今回の実験ではアニーリングスケジュールに関する検討を行っておらず，Tempered EM アルゴリズムにおける β は 1.0 とし，反復回数を 10 回に固定して実験を行った．

検索 API には Yahoo 検索 API [10]を使用した．検索 API では検索したい単語を組み込んだ URL を作成し，リクエストを送信することで，検索ヒット数（その単語が 1 回以上出現したページの数）やヒットしたサイトの URL 等の情報が xml 形式で返却される．その中の検索ヒット数を抜き出し，idf の計算に用いている．また，今回の実験では idf の計算に必要な全文書数を Yahoo が持つインデックス数である 192 億として計算している．

評価に用いる指標として補正パープレキシティを使用した．補正パープレキシティは式(16)によって求めることができる．

$$APP = \{ P(w_1 \cdots w_n) \cdot m^{-o} \}^{-\frac{1}{n}} \quad (16)$$

$P(w_1 \cdots w_n)$：単語列 $w_1 \cdots w_n$ が生成される確率

O：未知語の数

m：未知語の種類

Vol.2011-SLP-85 No.14
2011/2/5

5.2 実験結果と考察

実験結果を図 2 に示す．横軸に評価用の 15 講演を並べており，縦軸は各講演データにおける補正パープレキシティを計算している．

図 2 から分かるように，講演データにおいて差はあるが，15 講演すべてのデータにおいてパープレキシティの改善が見られた．また，通常の PLSA 言語モデルのパープレキシティをベースラインとすると，最大で約 35%，平均では約 26%のパープレキシティの削減効果が得られた．

分散分析表を表 1 に示す．群内の自由度が 28 で群間の自由度が 1 の時，1%有意水準で，F＝7.64 が棄却域の境目である．実験から得られた F 値は 38.9 であり，帰無仮説「通常 PLSA 言語モデルと提案手法のパープレキシティによる評価の平均に差はない」は 1%の有意水準で棄却され，通常 PLSA 言語モデルと提案手法では評価の平均に差があると言える．すなわち，本手法はパープレキシティ削減において有効である．

以上の結果から，話題判定に単語出現頻度だけではなく文書間頻度の情報も加えることにより，PLSA 言語モデルにおけるトピックの誤判定をより抑えることができたと考えられる．

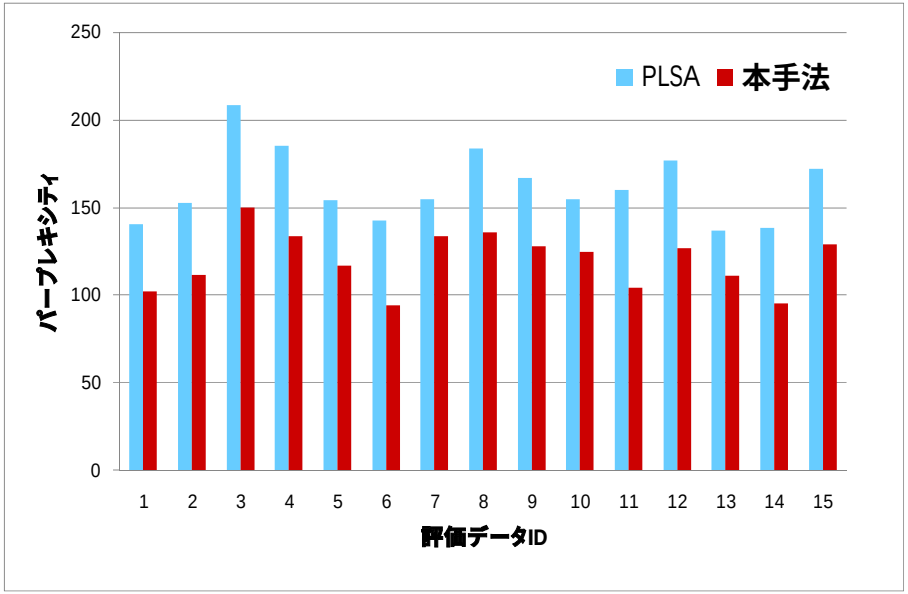

図 2：通常 PLSA 言語モデルと本手法のパープレキシティの比較

表 1：分散分析表

	平方和	自由度	平均平方	F値	P値
群間	13234.69	1	13234.69	38.96948	9.55253E-07
群内	9509.271	28	339.6168		
全体	22743.96	29			

6. まとめ

今回の報告では，トピックをモデル化する PLSA 言語モデルの改良について検討を行った．具体的には，PLSA 言語モデルにおけるトピック判定に用いる情報を単語出現頻度（tf）だけでなく，WWW から取得した idf も用いるという手法であり，tfidf の概念に基づいている．

この手法の有効性を検証するため，テストセットパープレキシティの比較によって，通常の PLSA 言語モデルとの性能比較を行った．

その結果，評価用に用いたすべてのデータにおいて通常の PLSA 言語モデルよりパープレキシティの改善が見られ，最大で約 35%の削減率に達した．このことから，tfidf の概念に基づいた PLSA 言語モデルにおける話題判定は有効であると考えられる．

参考文献

1) 北研二 ： 確率的言語モデル，東京大学出版会(1999)
2) 梶浦泰智，鈴木基之，伊藤彰則，牧野正三 ：WWW を利用した言語モデル教師なしタスク適応における有効検索クエリ決定法，情報処理学会研究報告, 2006-SLP-64
3) 増村亮，伊藤仁，伊藤彰則，牧野正三 ：WWW を利用した言語モデル適応のための検索クエリ構成の検討，情報処理学会研究報告, Vol.2009-SLP-76 No.10
4) 川端豪 ： 確率文法と話題マルコフモデルに基づく音声認識のための話題制御，電子情報通信学会論文誌, Vol.J77-D-Ⅱ, No.10, pp.1967-1972(1994)
5) 政瀧浩和，匂坂芳典，久木和也，河原達也 ：MAP 推定を用いた N-gram 言語モデルのタスク適応, SP96-103(1997)
6) Thomas Hofman : Plrobabilistic Latent Semantic Analysis, Uncertainity in Artificial Intelligence(1999)
7) Daniel Gildea, Thomas Hofmann : TOPIC-BASED LANGUAGE MODELS USING EM, EuroSpeech’99, pp.2167-2170
8) 栗山直人，鈴木基之，伊藤彰則，牧野正三 ：PLSA 言語モデルの学習最適化と語彙分割に関する検討，情報処理学会研究報告, 2006-SLP-60
9) 徳永健伸 ： 情報検索と言語処理，東京大学出版会(1999)
10) Yahoo 検索 API
http://developer.yahoo.co.jp/webapi/search/websearch/v2/websearch.html