LF-003

Skew Divergence に基づく母語話者/非母語話者文書の判別
Discrimination of Native/Non-native Documents Based on Skew Divergence

藤井　宏 † Hiroshi Fujii　田中　省作 ‡ Shosaku Tanaka　冨浦　洋一 § Yoichi Tomiura

1. はじめに

学習データから正しい言語知識を獲得したいと考える場合，その抽出源となる文書としては，文の流れが良く，文法の間違いや共起誤りなどをほとんど含まないようなものが望ましい．母語話者の記述した文書はこのような性質を有していると期待できる．よって，良質な学習データを大量に集めるための技術として，母語話者/非母語話者文書 (以降 N/NN と書く) の判別が重要になる．

この N/NN 判別法として，本稿では，判別対象の文書から品詞 n-gram(実際には n=3;trigram) を計算し，その分布が N/NN 文書における n-gram 分布[1] のいずれに近いかによって対象文書のクラスを決定する．確率分布間の相違度を測る尺度として，ゼロ頻度部分についても自然に取り扱える Skew Divergence[1][2] を用いる．さらにこの手法を改良し，個別文書における n-gram 分布のばらつきを考慮した判別法についても述べる．判定対象文書を特に英語科学技術論文に限定した小規模実験の結果，いずれの手法においても高精度での文書の N/NN 判別に成功した．

2. 母語話者/非母語話者文書の判別法

文書 d の発生源クラス C' を求めるために，次のような手法を考える．

> 文書 d から品詞 trigram 確率分布を推定し，文書クラス $C \in \{N, NN\}$ の品詞 trigram 確率分布との相違度を比較して，文書 d の発生源クラス C' を特定する．

一般に確率分布間の相違度を比較する尺度として，次式に示す KL Divergence が良く知られている．

$$D(p \parallel q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

品詞 trigram では品詞 2 個組の各条件下で確率分布が与えられるため，各条件部における分布間の相違度を併せて考慮する必要がある．そこで，複数の条件付確率分布を考慮した評価関数として，文書 d と文書クラス $C \in \{N, NN\}$ との相違度 $ED(d;C)$ を次のように定義する．

$$ED(d;C) = \sum_{ab \in Cond(d)} f_d(ab) D(P_d(\cdot|ab) \parallel P_C(\cdot|ab)) \quad (1)$$

ここで $Cond(d)$ は文書 d 中に出現する品詞 2 個組 ab の集合，$f_d(ab)$ は文書 d 中の品詞 2 個組 ab の出現頻度，$P_d(\cdot|ab)$ は文書 d から推定した条件 ab における trigram 確率分布，$P_C(\cdot|ab)$ はクラス C での条件 ab における真の trigram 確率分布とする．この相違度 ED を用いて，文書 d の発生源クラス C' は $C' = \operatorname{argmin}_C ED(d;C)$ で決定される．

式 (1) は次のように展開できる．

$$\begin{aligned} ED(d;C) &= \sum_{ab \in Cond(d)} f_d(ab) H(P_d(\cdot|ab)) \\ &\quad - \log \prod_{abc \in Tri(d)} (P_C(c|ab))^{f_d(abc)} \end{aligned}$$

ここで，$H(P)$ は確率分布 P のエントロピー，$Tri(d)$ は文書 d に現れる品詞 3 つ組の集合を表す．右辺第 1 項は文書 d のみに依存する値であり，第 2 項はクラス C から文書 d が発生する確率の対数で，$ED(d;C)$ はこの確率の減少関数になっている．したがって，相違度 ED を用いた判別法は各クラスからの文書 d の発生確率の比較に相当する．

前式ではクラス C における真の trigram 確率分布が既知の場合を想定したが，実際の問題に取り組む場合，一般に真の trigram 確率分布は未知である．クラス C の trigram 確率分布を推定する際，十分に信頼性の高い分布を見積もるためには膨大なデータ量を要する．しかし，クラスの確率分布を推定するために十分多くのデータを用意することは，N/NN 判別問題を解いて母語話者文書を収集する，という本来の目的に合わない．以降，文書クラス C の観測データから推定した条件 ab における trigram 分布を $\hat{P}_C(\cdot|ab)$ のように書くことにする．これを用いて $ED(d;C)$ を計算すると，次のような不具合が起こる．

特にクラス C の確率分布推定のための観測データが十分に大きくない場合，クラス C の観測データ中に一度も現れなかった 3 つ組が文書 d には現れることがある (C においてその発生確率が真にゼロなのか，データ不足で単に現れなかったのかは分からない；**ゼロ頻度問題**)．このような場合，推定された分布の信頼度に関わらず $ED(d;C) = \infty$ となり，判別に決定的な影響を与える．また，$f_d(x_1x_2x_3) > 0$ かつ $\hat{P}_N(x_3|x_1x_2) = 0$，$f_d(y_1y_2y_3) > 0$ かつ $\hat{P}_{NN}(y_3|y_1y_2) = 0$ のように，同一文書 d 内に N/NN どちらのクラスからの発生確率もゼロとなる部分が存在すると，発生源の判別が不可能になる．このような問題から，スムージングを施した推定確率を用いるアプローチが一般的である．しかしながら，スムージング手法はどれも理論的な根拠に乏しく，必ずしもそれが結果に良い影響を及ぼすとは限らない．

† 九州大学大学院システム情報科学府
‡ 九州大学情報基盤センター
§ 九州大学大学院システム情報科学研究院

[1] 本稿で n-gram 分布とは，特に注釈がない限り $n-1$ 個の品詞列が現れた条件の下で，その次にどのような品詞がどのくらいの確率で現れるかを表した条件付確率分布のことを指す．

そこで，相違度として KL Divergence 以外のものを用いることを考え，観測データ (学習データ) から推定した文書クラスの確率分布と対象文書 d との相違度として，どのような性質を有するものが望ましいかについて議論する．特に問題となるのは，文書 d 中での発生頻度が $f_d(x_1x_2x_3)(>0)$ 回で，クラス C において $\hat{P}_C(x_3|x_1x_2)=0$ となる $x_1x_2x_3$ が存在する場合であり，このとき $x_1x_2x_3$ における文書 d とクラス C との差異は次のような性質を満たしていることが望ましい．

1. C の真の確率分布において $P_C(x_3|x_1x_2)$ の値が大きくなるにつれ，推定確率 $\hat{P}_C(x_3|x_1x_2)=0$ となるような観測データが得られる可能性は小さくなる．また，d がクラス C から発生していると仮定すると，$f_d(x_1x_2x_3)/f_d(x_1x_2)(=P_d(x_3|x_1x_2))$ と $P_C(x_3|x_1x_2)$ は正の相関がある．したがって，$\hat{P}_C(x_3|x_1x_2)=0$ では $P_d(x_3|x_1x_2)$ の値が大きくなるほど，d と C との差異は大きくなるべきである．

2. C の真の確率分布において $P_C(x_3|x_1x_2)>0$ である場合，$\hat{P}_C$ を推定するために用いる観測データの量を増やすにしたがって，推定確率 $\hat{P}_C(x_3|x_1x_2)=0$ となるような観測データが得られる可能性は小さくなる．したがって，$f_d(x_1x_2x_3)>0$ ならば $\hat{P}_C(x_3|x_1x_2)=0$ を推定するために用いた観測データの量が大きいほど，d と C との差異は大きくなるべきである．

ここで，文献 [1] で提案されている Skew Divergence に注目する．Skew Divergence は，次式で定義される．

$$\begin{aligned} s_\alpha(p,q) &= D(p \parallel \alpha q+(1-\alpha)p) \\ &= \sum_{x\in X} p(x)\log\frac{p(x)}{\alpha q(x)+(1-\alpha)p(x)} \end{aligned}$$

ただし，α は $0\le\alpha\le1$ の定数．この Skew Divergence は，先ほど挙げた性質に綺麗に対応していることを示す．$p(x)=P_d(x|x_1x_2)$，$q(x)=\hat{P}_C(x|x_1x_2)$ とおくと，

- まず，$q(x)=0$ における差異は，$-p(x)\log(1-\alpha)=-P_d(x|x_1x_2)\log(1-\alpha)$ となり，これは $P_d(x|x_1x_2)$ に比例した増加関数である (性質 1).
- また，α の値を，分布 q を推定した際の観測データ量の増加関数で与えると，$-p(x)\log(1-\alpha)$ はその増加関数になる (性質 2).

加えて，$\alpha<1$ となるように α を選べば，Skew Divergence は必ず有限の値をとるため，KL Divergence を用いる場合のような問題は起きない．以上のことから，観測データから推定したクラスの確率分布と判別対象文書 d との相違度を測るのに Skew Divergence を用いることの有効性が示された．

式 (1) での $ED(d;C)$ の定義と同様にして，Skew Divergence を用いた文書 d と文書クラス C の相違度 $ES(d;C)$ を次のように定義する．

$$ES(d;C)=\sum_{ab\in Cond(d,N,NN)} f_d(ab)s_\alpha(P_d(\cdot|ab),\hat{P}_C(\cdot|ab)) \quad (2)$$

ここで，$Cond(d,N,NN)$ は文書 d，クラス N，クラス NN それぞれの分布を推定するために用いた観測データにおいて，共通に条件部として出現した品詞 2 個組の集合 $(=Cond(d)\cap Cond(N)\cap Cond(NN))$ である[2]．この相違度 ES を用いて，文書 d の発生源クラス C' は $C'=\operatorname{argmin}_C ES(d;C)$ で決定される．

3. 判別に用いる trigram 分布の選択指針

クラス C の各条件における trigram 分布を学習データから推定した際，中には判別に用いるのに不適当と思われる分布が含まれることがある．例えば，文書著者毎の個性が現れ，同一クラス内の各文書で trigram の出現傾向に大きな違いがあるような分布である．このようにクラス内個別文書間の共通性が低い分布は N/NN というクラス判別に有効に働くとは考えにくく，判別の際に用いる評価材料から除外したい．

上記のフィルタリングは，クラス全体から推定した分布を中心とした，クラス内の個々の文書における分布のばらつき具合を見ることで実現できる．クラス C 内における条件部 ab の trigram 分布のばらつき具合は，次に定義する $MD(ab;C)$ で計算する．

$$MD(ab;C)=\frac{1}{|Ds(ab;C)|}\sum_{d\in Ds(ab;C)} s_\alpha(P_d(\cdot|ab),\hat{P}_C(\cdot|ab))$$

ここで $Ds(ab;C)$ はクラス C の学習データに含まれる文書のうち，条件部 ab を持つ trigram が出現する文書の集合である．$MD(ab;C)$ はクラス C の文書全体から推定した条件 ab における分布と，同じ条件部を持ったクラス C 内の個別の文書から推定した分布との Skew Divergence の平均になっている（これを平均偏差と呼ぶ）．これを用いて，各条件部について trigram 分布のクラス C 内でのばらつきを見積もり，平均偏差がある閾値以上大きい分布に関しては判別の際の評価材料から除外する．

また，$Ds(ab;C)$ に含まれる文書数が非常に小さい場合，その分布はクラス C 内でもごく一部の文書にしか現れないことを表している．このような分布もクラス C 全体が共通して持つ傾向とは考えられないため，評価材料から除くべきである．

これら 2 つの基準によるフィルタリングにより，判別の際の評価に用いる trigram 分布の条件部を次式で制限する．

$$RCond(C)=\{\ ab \mid MD(ab;C)<\theta, |Ds(ab;C)|>\beta\}$$

ここで θ,β は定数．この制限を式 (2) に適用した相違度として，$RS(d;C)$ を次式のように定義する．

$$RS(d;C)=\sum_{ab\in R(d)} f_d(ab)s_\alpha(P_d(\cdot|ab),\hat{P}_C(\cdot|ab)) \quad (3)$$

ここで，$R(d)=Cond(d)\cap RCond(N)\cap RCond(NN)$．この相違度 $RS(d;C)$ を $ES(d;C)$ の代わりに用いることで，N/NN 判別性能の向上が期待できる．

[2]観測データ内に出現しなかった条件部についての trigram 分布は推定不能のため，このような条件部は判別の評価材料から除く．

表 1: N/NN 判別実験結果 (数値は正解率)

Method	Closed Test	Open Test
Baseline	0.784	0.784
ExSkew, $\alpha = 0.80$	0.997	0.969
ReSkew, $\alpha = 0.90$	0.989	0.979

4. 実験

4.1 データと方法

評価実験について述べる．データは，母語話者文書集合として 229 文書，非母語話者文書集合として 63 文書を用意し[3]，予め Tree Tagger [4] によって品詞列に変換しておく．実験は，教師データとテストデータが同一のクローズド実験と，4-交差検定[5] のオープン実験の 2 通りを行い，その正解率で評価する．ここで正解率とは，テストデータ中の文書が本来のクラスに判別された割合である．なお Skew Divergence の定数 α は $0.5 \leq \alpha < 1$ を適当に動かし，オープン正解率が最良のものを示している．

4.2 結果

N/NN 判別実験の結果を表 1 に示す．なお，Baseline は N/NN 文書比で，常に N と判別した場合の正解率である．ExSkew は式 (2) の *ES*，ReSkew は式 (3) の *RS* を用いた判別法である．この結果から，両提案手法はクローズド/オープン正解率ともに大幅にベースラインを超えており，これは検定によると有意水準 1%で有意な差があり，提案手法の有効性が確認された．ExSkew と ReSkew の正解率の差異は，有意なものではなかった．なお，いずれの手法でも Skew Divergence のパラメタ α の値が正解率に大きく影響する．

5. 関連研究

N/NN 判別問題の従来研究として，文献 [3] では，判定文書の発生確率を N/NN の 2 つの n-gram モデルを用いて推定し，より発生確率が高いモデルをその文書のクラスとして決定する．文献 [3] の手法はゼロ頻度問題を回避するために，n-gram に対しバックオフや線形補完といったスムージングを行うことを前提としているが，スムージングは理論的な根拠に乏しく，必ずしも判定に良い結果を与えるわけではない．

また，本研究に関連した話題として文書の著者判別が挙げられる．文献 [4] では文字 n-gram の同時確率分布を用いた著者判別法が提案されており，アプローチは我々の提案手法と類似しているが，ゼロ頻度部分の取り扱いなどが大きく異なる (文献 [4] では確率ゼロの n 個組を分布の比較対象から除外している)．また，著者判別は文書における個人の癖に注目する問題で，N/NN 判別では複数の著者が同じクラスに混在するため，判別に利く要因が異なることが予想される．

6. おわりに

本稿では，文書とクラスの品詞 n-gram 分布を Skew Divergence で比較して文書の N/NN を判別する手法を提案し，その有効性を示した．

今後，より大規模な実験，Skew Divergence の α やフィルタリングの閾値など実験的に求めたパラメタの理論的な求め方に関する研究，さらに N/NN の判別に有用な trigram 分布の分析，吟味を予定している．特に最後の課題については従来専門家の内省では考えられないような母語話者性に関連する要因が含まれている可能性があり，文書の判別のみならず，英語教育，第二言語習得への成果の還元が期待される．また，式 (1) においてバックオフや線形補間などのスムージング手法を用いた場合との比較実験，ナイーブベイズなど一般に知られる文書分類手法との性能比較も行う予定である．

参考文献

[1] Lillian Lee.:Measures of Distributional Similarity, *Proceedings of the 37th ACL*, pp.25–32 (1999).

[2] Lillian Lee.:On the Effectiveness of the Skew Divergence for Statistical Language Analysis, *Artificial Intelligence and Statistics 2001*, pp.65–72 (2001).

[3] 緒方 伸輔, 田中 省作, 冨浦 洋一.:非内容語の n-gram に基づく英語母語話者性の推定, 情報処理学会研究報告 2004-NL-160, (2004).

[4] 松浦 司, 金田 康正.:近代日本小説家 8 人による文書の n-gram 分布を用いた著者判別, 情報処理学会研究会報告 2000-NL-137, (2000).

[3]母語話者文書として海外の国際会議で発表された論文で著者の所属機関が USA のもの，非母語話者文書として日本の国際会議で発表された論文で著者が日本人であるものを使用した．

[4]`http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`

[5]N/NN データをランダムに 4 分割し，3 つを教師データ，残る 1 つをテストデータとし正解率を計算する．これを順番に 4 回繰り返し，平均正解率を求める．