中近世スペイン語古文書の統計的年代推定・場所推定

川崎 義史
上智大学（日本学術振興会特別研究員 PD）
kyossii@gmail.com

1. はじめに

本研究の目的は、**中近世スペイン語古文書**（こもんじょ）の作成年代と作成場所を統計的に推定する方法を開発することである。古文書とは、法令、権利、財産譲渡などの事項を明記した近代以前の手書きの行政・法律関係書類の総称である。現存する文献史料に基づく歴史学や通時言語学の研究において、作成年代と作成場所が不明の文献がいつ・どこで作成されたかを特定すること、及び文書の真贋を判定することは最重要課題である。本研究では、古文書を文書、文書の作成年代の推定を**年代推定**、文書の作成場所の推定を**場所推定**と呼ぶことにする。

データセットとして、中近世スペイン語古文書コーパス **CODEA+ 2015**（Corpus de Documentos Españoles Anteriores a 1800）を使用した。このコーパスは、スペインのアルカラ大学の文献学研究グループ GITHE（Grupo de Investigación de Textos para la Historia del Español）が作成しているもので、1100 年から 1700 年の間に、主にスペインで作成された 1500 の古文書からなる。このうち 1237 文書は、作成年代と作成場所が一義的に決まる。

2. 関連研究

年代を離散変数とみなし、年代推定を文書分類の枠組みで扱った研究としては、De Jong *et al.* (2005)、Tilahun *et al.*（2012)、川崎（2015）等が挙げられる。同様に、地点を離散変数とみなし、場所推定を文書分類の枠組みで扱った研究としては、Cheng *et al.*（2010）、Eisenstein *et al.*（2010）、Roller *et al.*（2012）、Han *et al.*（2014）、Wing & Baldridge（2014）、Hulden *et al.*（2015）等が挙げられる。

上記の年代推定の研究では言語の空間的変異を無視している。同様に、場所推定の研究では言語の時間的変異を無視している。しかし、同年代における空間的変異や同地域における時間的変異が存在するので、言語の変異を扱う際には、時間軸と空間軸を同時に考慮する必要がある。

3. 問題定義

本研究は、年代推定・場所推定を文書分類の問題として考える。文書分類は、属するクラスが不明の文書を、既定のクラスのいずれかに分類するタスクである。分類するクラスは文書の作成年代と作成場所であり、どちらも離散変数とみなす。粒度を 1 年とした 1100 年から 1700 年の間の年代tの集合を$T \equiv \{1100, 1101, \dots, 1700\}$とする（$|T| = 601$）。また、粒度を現代スペインの自治州とした地点$l$の集合を$L \equiv \{\mathrm{AN}, \mathrm{AR}, \dots, \mathrm{VA}\}$とする（$|L| = 17$）。紙幅の関係で、粒度を県とした場合（$|L| = 42$）の分析は省略する。

推定は、作成年代と作成場所を個別に推定する**個別推定**と、両者を同時に推定する**同時推定**の二つの方法で行う。後者の方法は、管見の限り、本研究が初めて提案する方法である。年代推定・場所推定を行う文書qの推定年代を$\hat{t_q}$、推定場所を$\hat{l_q}$とする。個別推定では、$|T|$個の年代から推定年代$\hat{t_q}$を、$|L|$個の地点から推定場所$\hat{l_q}$を独立に求める（3.1）。総クラス数は$|T| + |L|$となる。

$$\hat{t_q} = \arg\max_{t \in T} \phi(q, t)$$

$$\hat{l_q} = \arg\max_{l \in L} \phi(q, l) \tag{3.1}$$

ここで、$\phi(q, t)$は文書qと年代tとの類似度を、$\phi(q, l)$は文書qと地点lとの類似度を表す関数である。

一方、同時推定では、$|T| \times |L|$個のクラスから、作成年代と作成場所の組み合わせを同時に推定する。年代tと地点lの組み合わせを(t, l)、(t, l)の集合を(T, L)とする。このとき文書qの推定年代$\hat{t_q}$と推定場所l_qの組み合わせ$(\hat{t}_q, \hat{l}_q)$は、$\phi(q, (t, l))$が最大となるクラスになる（3.2）。$\phi(q, (t, l))$は文書qと(t, l)の類似度を表す関数である。

$$(\hat{t}_q, \hat{l}_q) = \arg\max_{(t,l) \in (T,L)} \phi(q, (t, l)) \tag{3.2}$$

同時推定のメリットは、同一年代における空間的変異や同一地域における時間的変異を考慮することができる点である。これにより、個別推定に比べ、推定精度が向上すると期待される。デメリットは、個別推定に比べクラス数が多くなるので、計算量が増加する点である。

本研究では、文書qとクラスcとの類似度を表す関数$\phi(q, c)$として *n*-gram 言語モデル（3.2.節）と JS 情報量（3.3.節）を用いる。素性値としては、文字 *n*-gram のカーネル平滑化頻度を用いる（3.1.節）。大文字小文字は区別せず、スペースも一つの文字としてカウントする。文字 *n*-gram を素性とする理由は、現時点では中近世スペイン語をレン

マ化もしくはステミングする技術がないため、また古文書のように文書長が短い文書を用いる場合、文字以外に出現頻度の高い素性を抽出することが困難なためである。文字*n*-gram の抽出に先立ち、前処理として、アルファベット以外の文字の削除、文書中の明示的な作成年代や作成場所の削除、本文のメタ情報の削除等を行った。

3.1. カーネル平滑化

カーネル平滑化（kernel smoothing）とは、カーネル関数を用いて、関数$f(x)$からより滑らかな関数$\hat{f}(x)$を推定する手法である（Hastie *et al.* 2009: Chapter 6）。本研究において、$f(x)$は、各クラスにおける文字 *n*-gram の出現頻度に当たる。カーネル平滑化により、データセットに点在する欠損値の補充、スパースネスの緩和、頑健な推定が可能になる。

着目年代t'に対する年代tの重みを表す時間カーネル関数$K(t,t')$を、ガウスカーネルを用いて定義する（3.3）。σ_tは時間カーネル平滑化パラメータである。

$$K(t,t') = \frac{1}{\sqrt{2\pi}\sigma_t}\exp\left(-\frac{\|t-t'\|^2}{2{\sigma_t}^2}\right) \tag{3.3}$$

ここで、ある素性の年代 tにおける出現頻度をn_tとすると、年代 t'におけるカーネル平滑化頻度$\hat{n}_{t'}$は、局所重み付き平均として表される（3.4）。地点の情報は無視している。

$$\hat{n}_{t'} = \frac{\sum_{t\in T} K(t,t') * n_t}{\sum_{t\in T} K(t,t')} \tag{3.4}$$

計算量を抑えるために、$|t-t'| \leq 20$となるような年代tのみを考慮する。

同様に、着目地点l'に対する地点lの重みを表す空間カーネル関数$K(l,l')$を、ガウスカーネルを用いて定義する（3.5）。σ_lは空間カーネル平滑化パラメータである。

$$K(l,l') = \frac{1}{\sqrt{2\pi}\sigma_l}\exp\left(-\frac{\|l-l'\|^2}{2{\sigma_l}^2}\right) \tag{3.5}$$

二地点間の距離$\|l-l'\|$は Google Maps の徒歩距離で近似した[1]。距離の単位は㎞である。ここで、ある素性の地点lにおける出現頻度をn_lとすると、地点l'におけるカーネル平滑化頻度$\hat{n}_{l'}$は、局所重み付き平均として表される（3.6）。年代の情報は無視している。

$$\hat{n}_{l'} = \frac{\sum_{l\in L} K(l,l') * n_l}{\sum_{l\in L} K(l,l')} \tag{3.6}$$

計算量を抑えるために、$|l-l'| \leq 200$となるような地点lのみを考慮する。

本稿は、Tilahun *et al.*（2012）の *r* 次元カーネル（*r*-dimensional kernel）に基づき、時間と空間の両方を考慮した二次元の**時空間カーネル平滑化**を提案する。時間軸と空間軸の各々においてカーネル平滑化を行う先行研究は存在するが（Cheng *et al.* 2010; Tilahun *et al.* 2012; Hulden *et al.* 2015 等）、両者を組み合わせた時空間カーネル平滑化を提案した研究は管見の限り存在しない。

ある素性の年代 t、地点lにおける頻度を$n_{t,l}$とする。このとき、年代 t'、地点l'におけるカーネル平滑化頻度$\hat{n}_{t',l'}$は、局所重み付き平均として表される（3.7）。年代tと地点lには相関がないとする。

$$\hat{n}_{t',l'} = \frac{\sum_{t\in T}\sum_{l\in L} K(t,t') * K(l,l') * n_{t,l}}{\sum_{t\in T}\sum_{l\in L} K(t,t') * K(l,l')} \tag{3.7}$$

時間カーネル関数$K(t,t')$と空間カーネル関数$K(l,l')$の積からなる$K(t,t') * K(l,l')$を、時空間カーネル関数と呼ぶことにする。σ_tとσ_lは、年代、地点、素性には依存せず全クラスで一定とする。

3.2. *n*-gram 言語モデルによる文書分類

n-gram 言語モデル（*n*-gram language model）は、文書の生成確率（尤度）をモデル化した確率的言語モデルである（Chen & Goodman 1998）。長さlの文字列$w_1^l = w_1 w_2 \dots w_l$として表される文書$q$がクラス$c$において生成される確率$P(q|c)$を（3.8）のようにモデル化する。

$$P(q|c) = \prod_{i=1}^{l+1} P_{c:AD}\left(w_i \middle| w_{i-n+1}^{i-1}\right) \tag{3.8}$$

ここで、w_0は文頭記号、w_{l+1}は文末記号である。$P_{c:AD}\left(w_i \middle| w_{i-n+1}^{i-1}\right)$は、訓練データから加算スムージングにより求めた、クラスcにおける文字w_iの条件付き確率の推定値である（3.9）。

$$P_{c:AD}\left(w_i \middle| w_{i-n+1}^{i-1}\right) = \frac{n\left(w_{i-n+1}^{i-1}w_i, c\right) + \alpha}{\sum_{w\in V} n\left(w_{i-n+1}^{i-1}w, c\right) + \alpha|V|} \tag{3.9}$$

ここで、$n(w,c)$はクラスcにおける文字列wの頻度、Vはデータセット全体の文字集合、$|V|$は文字の種類数、$\alpha \in (0,1]$は加算スムージングパラメータである。

文書qの属するクラス$\hat{c_q}$は、対数尤度$\log P(q|c)$が最大となるクラスcとする（3.10）。Cはクラスの集合である。

$$\hat{c_q} = \arg\max_{c\in C} \sum_{i=1}^{l+1} \log P_{c:AD}\left(w_i \middle| w_{i-n+1}^{i-1}\right) \tag{3.10}$$

3.3. JS 情報量による文書分類

JS 情報量（Jensen-Shanon divergence）は、同じ事象空間上の二つの確率分布間の差異を情報理論に基づき測る尺度である（高村 2010 1.6）。ある事象空間における文書qの確率分布をP_q、クラスcの確率分布をP_cとすると、二つの確率分布の JS 情報量$D_{JS}(P_q||P_c)$は（3.11）で与えられる。

$$D_{JS}(P_q||P_c) = \frac{D_{KL}(P_q||R) + D_{KL}(P_c||R)}{2} \tag{3.11}$$

ここで、$D_{KL}(\cdot||\cdot)$は二つの確率分布の KL 情報量、Rは二

[1] https://www.google.co.jp/maps（2015 年 5 月 21 日アクセス）。

つの確率分布P_qとP_cの平均として定義される確率分布である。確率分布の値は最尤推定で求めた。

JS 情報量が小さいほど、二つの確率分布の類似度は高い。よって、文書qの属するクラス$\widehat{c_q}$は、JS 情報量の逆数$D_{JS}(P_q||P_c)^{-1}$を用いて、(3.12) のように予測される。

$$\widehat{c_q} = \operatorname*{arg\,max}_{c \in C} D_{JS}(P_q||P_c)^{-1} \tag{3.12}$$

4. 実験

4.1. 実験設定

データセットは、作成年代と作成場所が既知の 1237 文書である。データセットを約 10 対 1 で訓練データとテストデータに分割し、10 分割交差検定を行った。

本研究では、テストデータに属する文書qの作成年代t_qと作成場所l_qがともに不明だと仮定し、推定年代$\widehat{t_q}$と推定場所$\widehat{l_q}$を求める実験を行った。推定方法は、個別推定と同時推定の二種類である。個別推定では、作成年代と作成場所を独立に推定する。同時推定では、両者を同時に推定する。素性には、文字 2-gram を使用した。

最適なパラメータは、グリッドサーチで求めた。予備実験の結果を踏まえ、パラメータ空間は、時間カーネル平滑化パラメータ$\sigma_t \in \{\delta, 3, 5, 10\}$、空間カーネル平滑化パラメータ$\sigma_l \in \{\delta, 25, 50, 75, 100\}$、加算スムージングパラメータ$\alpha \in \{0.001, 0.01, 0.1, 1\}$とした。時間カーネル平滑化パラメータと空間カーネル平滑化パラメータにおけるδはほぼゼロと見なせる正の値 ($\delta \approx +0$) で、カーネル平滑化しないことを意味している。実験は、すべて筆者が Excel VBA で実装して行った。

年代推定・場所推定の評価指標は、いずれも絶対値誤差平均 (Mean Absolute Error: 以下 MAE)、二乗平均平方根誤差 (Root Mean Squared Error: 以下 RMSE)、絶対値誤差中央値 (Median Absolute Error: 以下 MedAE) である。年代推定の絶対値誤差平均をMAE_t、二乗平均平方根誤差をRMSE_t、絶対値誤差中央値をMedAE_t、場所推定の絶対値誤差平均をMAE_l、二乗平均平方根誤差をRMSE_l、絶対値誤差中央値をMedAE_lとする。

年代推定・場所推定の結果をまとめて評価するために、RMSE_tとRMSE_lの積で与えられる評価指標を定義する (4.1)。

$$\mathrm{STEP} = \mathrm{RMSE}_t * \mathrm{RMSE}_l \tag{4.1}$$

これを時空間誤差積 (Spatio-Temporal Error Product: 以下 STEP) と呼ぶ。STEP が小さいほど、推定精度が高いと評価する。STEP の単位は年㎞である。

本研究の提案する同時推定と時空間カーネル平滑化の有効性を確認するために、*n*-gram 言語モデルと JS 情報量による分類のそれぞれにおいて、推定方法の違い (個別推定か同時推定か) とカーネル平滑化の有無による STEP の値を比較する。組み合わせは、カーネル平滑化なしの個別推定、カーネル平滑化なしの同時推定、時間カーネル平滑化・空間カーネル平滑化ありの個別推定、時空間カーネル平滑化ありの同時推定の 4 パターンである。各パターンにおいて STEP が最小となるモデルを比較し、時空間カーネル平滑化ありの同時推定の STEP が最小となれば、本研究の提案する手法の有効性が確認される。

4.2. 結果

表 1 に、各パターンにおいて STEP が最小となるモデルを示す。言語モデルは *n*-gram 言語モデルのことである。

5. 考察

STEP が最小となるのは、*n*-gram 言語モデルでも JS 情報量による推定でも、時空間カーネル平滑化ありの同時推定の場合である。したがって、本研究の提案する手法の有効性が確認された。次に STEP が小さいのは、カーネル平滑化なしの同時推定の場合である。その次に STEP が小さいのは、*n*-gram 言語モデルでは時間カーネル平滑化・空間カーネル平滑化ありの個別推定、JS 情報量による推定ではカーネル平滑化なしの個別推定である。STEP が最大となるのは、*n*-gram 言語モデルではカーネル平滑化なしの個別推定、JS 情報量による推定では時間カーネル平滑化・空間カーネル平滑化ありの個別推定である。

n-gram 言語モデルによる推定では、カーネル平滑化の有無にかかわらず、個別推定よりも同時推定の推定精度が高くなる。また個別推定でも同時推定でも、カーネル平滑化ありの方が、カーネル平滑化なしの場合より推定精度が高くなる。JS 情報量による推定でも、カーネル平滑化の有無にかかわらず、個別推定よりも同時推定の推定精度が高くなる。同時推定では、カーネル平滑化ありの方が、カーネル平滑化なしの場合より推定精度は高くなる。しかし、個別推定の場合、カーネル平滑化を行うと、推定精度が低下してしまう。

カーネル平滑化を行う場合、時間カーネル平滑化パラメータσ_tは小さい値 ($\sigma_t = 3$) が効果的なようである。空間カーネル平滑化のパラメータσ_lについては、特に効果的な値は見られない。ただし$\sigma_l = 25$は、事実上、空間カーネル平滑化を行っていないことになるので、空間カーネル平滑化の効果は限定的な可能性がある。加算スムージングのパラメータαは 1 より小さい値が効果的なようである。

分類	推定方法	σ_t	σ_l	α	年代推定			場所推定（自治州）			STEP
					MAE_t	RMSE_t	MedAE_t	MAE_l	RMSE_l	MedAE_l	
言語モデル	個別	δ	δ	0.001	18.63	31.70	10.00	193.88	330.73	152.00	10485.43
言語モデル	同時	δ	δ	0.1	16.28	30.65	8.00	121.14	248.24	0.00	7607.50
言語モデル	個別	3	25	0.001	17.87	30.41	10.00	194.19	330.52	174.00	10051.19
言語モデル	**同時**	**3**	**100**	**0.001**	**14.22**	**24.19**	**8.00**	**141.63**	**237.02**	**73.00**	**5734.00**
JS 情報量	個別	δ	δ	*	16.48	27.35	9.00	333.90	551.95	242.00	15094.94
JS 情報量	同時	δ	δ	*	15.68	27.36	8.00	121.81	257.79	0.00	7053.58
JS 情報量	個別	3	25	*	20.21	33.01	12.00	335.25	549.19	242.00	18129.16
JS 情報量	**同時**	**3**	**75**	*	**15.05**	**25.12**	**9.00**	**132.79**	**234.77**	**0.00**	**5896.19**

表 1　推定方法の違いとカーネル平滑化の有無による年代推定・場所推定（自治州）の最良の推定精度

6.　おわりに

本稿では、中近世スペイン語古文書の作成年代と作成場所を統計的に推定する方法を提案した。中近世スペイン語古文書コーパス CODEA+2015 を用いた実験の結果、適切なパラメータの下で、本稿で提案した作成年代と作成場所の同時推定が、個別推定に比べ、予測精度が高くなること、また、本稿で提案した時空間カーネル平滑化を行った方が、行わない場合に比べ、推定精度が高くなることを示した。ただし、そのトレードオフとして、計算量は増加する。

本研究では、文字 *n*-gram の出現頻度が文書の作成年代と作成場所のみに依存すると仮定し、文書の内容、作成者や発行機関、差出人、受取人等の属性は無視した。しかし、これらの社会言語学的要因も文字 *n*-gram の出現頻度に影響していると考えられる。したがって、これらの変数も考慮した分類手法の開発を今後の課題としたい。

謝辞

本研究は JSPS 科研費 15J04335 の助成を受けたものである。

参考文献

Chen, S., & Goodman, J. (1998). An empirical study of smoothing techniques for language modeling. *Technical Report TR-10-98, Harvard University*.

Cheng, Z., Caverlee, J., & Lee, K. (2010). You are where you tweet: A content-based approach to geo-locating Twitter users. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 759-768.

De Jong, F., Rode, H., & Hiemstra, D. (2005). Temporal language models for the disclosure of historical text. *Proceedings of the 16th International Conference of the Association for History and Computing*, 161-168.

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2010). A latent variable model for geographical lexical variation. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1277-1287.

GITHE (Grupo de Investigación Textos para la Historia del Español): *CODEA+ 2015 (Corpus de Documentos Españoles Anteriores a 1800)*. http://corpuscodea.es/

Han, B., Cook, P., & Baldwin, T. (2014). Text-based Twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49, 451-500.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Second ed.). Springer New York.

Hulden, M., Silfverberg, M., & Francom, J. (2015). Kernel density estimation for text-based geolocation. *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 145-150.

Roller, S., Speriosu, M., Rallapalli, S., Wing, B., & Baldridge, J. (2012). Supervised text-based geolocation using language models on an adaptative grid. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1500-1510.

Tilahun, G., Feuerverger, A., & Gervers, M. (2012). Dating medieval English charters. *The Annnals of Applied Statistics*, 6(4), 1615-1640.

Wing, B., & Baldridge, J. (2014). Hierarchical discriminative classification for text-based geolocation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 336-348.

川崎義史. (2015). 機械学習による中世スペイン語古文書の作成年代推定法. 言語処理学会第21回年次大会発表論文集, 333-336.

高村大也. (2010). 言語処理のための機械学習入門（奥村学監修）. 東京: コロナ社.