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

川崎 義史
東京大学大学院　総合文化研究科（日本学術振興会特別研究員 PD）
kyossii@gmail.com

1. はじめに

本論文では、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を提案する。ここで古文書（こもんじょ）とは、財産譲渡や権利関係などに関する行政・法律文書の総称である。現存する文献史料に基づく歴史学・歴史言語学の研究において、作成年代不詳文書の年代推定（dating）及び文書の真贋判定（verification）は最重要課題であり、作成年代推定法の開発には大きな意義がある。

2. 関連研究

スペイン語文献学では、Azofra（2009: 201-204）が、言語的特徴に基づいた年代推定法を提案した。各年代は形態統語論的特徴の二値ベクトルで表現され、推定年代は論理式により導かれる。ただし、年代区分が世紀ごとであり正確な年代推定はできない。

Tilahun *et al.*（2012）は、中世イングランド（1089~1438年）で発行されたラテン語勅許状（charter）の年代推定法を開発した。単語 N-gram による言語モデルを構築し、尤度が最大となる年代を推定年代とした。約 3400 の文書を用いた実験の結果、絶対値誤差平均は 9 年となった。

この他、文書分類（Text Categorization）や著者推定（Author Attribution）とも、分類すべき属性（トピック、著者）は異なるが、方法論では共通する点が多い。

3. コーパス

本研究では、中世スペイン語古文書コーパス CODEA（Corpus de Documentos Españoles Anteriores a 1700）を使用した。これは、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）が作成しているコーパスで、1100 年から 1700 年の間にスペイン各地で発行された約 1500 の古文書からなる。このうち、約 100 文書の作成年代が不明である。現時点では、コーパスに品詞や語幹等の情報は一切付加されていない。

単語数で測った文書長は裾が重い分布となっている（図 1）。平均文書長は 697 語、中央値は 497 語、最小値は 51 語、最大値は 6735 語である。総単語数は約 105 万語である。文書数も年代毎に大きく異なる（図 2）。

図 1　文書長別（横軸）の文書頻度（縦軸）

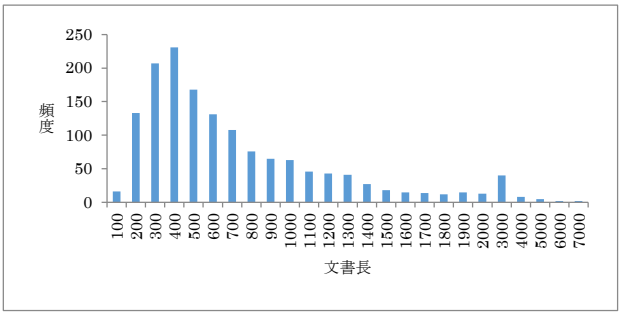
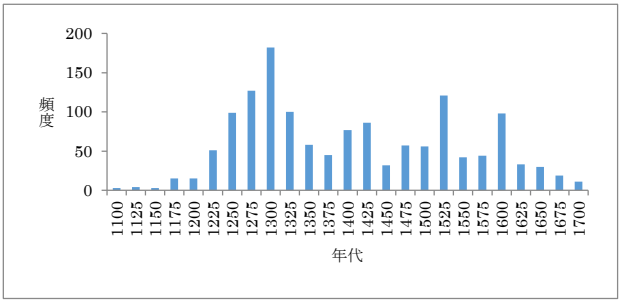


図 2　年代別（横軸）の文書頻度（縦軸）



以下に ID 1 の文書（1251 年）の最初の数行を示す。<…>は省略記号を復元した部分である。

> Connoscida cosa sea a todos los q<ue> esta carta uieren como yo don Fferrando por la gra<cia> de dios Rey de Castiella de Toledo de Leon de Gallizia de Seuilla de Cordoua de Murc<ia> & de Jahen enbie mis cartas a uos el Conceio de Guadalfaiara q<ue> enbiassedes u<uest>ros om<n>es buenos de u<uest>ro Conceio a mj …

4. 素性（Feature）

本研究では、2 種類の素性セットを用いる。一つ目は、スペイン語文献学の知見（Azofra 2009; Menéndez Pidal

1999; Penny 2002 等）に基づいた約 300 の言語的特徴（F1）である。これらの変数は、通時変異・地域変異・文書種類変異（勅令、教会関係、裁判文書等）を表すものである。たとえば、動詞 tener（英語の maintain や contain などの-tain に相当し、「持つ」の意味）の点過去（過去形の一種）には、tove と tuve の二つの変異形が存在し、その分布は年代推定の手掛かりとなる。素性セット F1 は、後述の *k*-近傍法とナイーブベイズ分類器で用いる。

もう一つの素性セットは、文字 2-gram（F2）である。文字 N-gram を選択したのは、単語 N-gram に比べ、素性数を大幅に削減できるからである。中世スペイン語には、英語で使用されるアルファベット 26 文字に加え、áéíóúñç や省略記号等が用いられるため、約 35 の文字が存在する。文頭・文末記号の挿入、大文字の小文字への変換、文書中の作成年代（たとえば M CC Lxxx Nona）の削除、省略箇所<…>の「@」への置換を行った結果、2-gram の総数は約 900 となった。文字 N-gram の通時変化は、文献学的には、言語の音素配列（phonotactics）もしくは文字配列（graphotactics）の通時変化であると考えることができる（たとえば、/i/を表す<i~j~y>の出現位置の通時変化）。素性セット F2 は、後述の N-gram モデルと *k*-近傍法で用いる。素性選択は行っていない。

5. 平滑化（Smoothing）

欠損年代の補完、スパースネスの緩和、頑健な推定を行うために、離散変数である 1 年刻みの年代を連続変数に読み替え、素性の出現頻度の平滑化を行った。$t \in [1100,1700]$を年代、$n_{f,t}$を素性fの年代tにおける出現頻度とし、平滑化後の出現頻度$n'_{f,t}$をガウスカーネル$K(t,t')$を用いて以下のように定義する。

$$n'_{f,t} = \frac{1}{Z(t)} \sum_{t'=1100}^{1700} K(t,t') * n_{f,t}$$

$$= \frac{1}{Z(t)} \sum_{t'=1100}^{1700} exp\left(-\left(\frac{t-t'}{\sigma}\right)^2\right) * n_{f,t}$$

ここで、$Z(t)$は重みの和を 1 とするための正規化定数である。

$$Z(t) = \sum_{t'=1100}^{1700} K(t,t') = \sum_{t'=1100}^{1700} exp\left(-\left(\frac{t-t'}{\sigma}\right)^2\right)$$

$\sigma \in \{5,10\}$で平滑化した出現頻度は、後述の N-gram モデルとナイーブベイズ分類器で用いる。ただし、N-gram モデルでは小数点以下を切り上げて整数値にした。ナイーブベイズ分類器では、年代tの文書数N_tにも同様の平滑化を行った。

6. 分類器（Classifier）

6.1. N-gram モデル（N-gram model）

Kneser-Ney 法を用いた 2-gram の確率$P_{KN}(w_i|w_{i-1})$は下式で与えられる（Chen and Goodman 1998; Jurafsky and Manning）。

$$P_{KN}(w_i|w_{i-1}) = \frac{\max(c(w_{i-1},w_i) - d, 0)}{c(w_{i-1})} + \lambda(w_{i-1}) P_{CONTINUATION}(w_i)$$

$c(w_{i-1},w_i)$、$c(w_i)$はそれぞれ文字列$w_{i-1}w_i$、w_iの出現頻度である。割引値dは$d = n_1/(n_1 + 2n_2)$で与えられ、n_1、n_2はそれぞれ出現頻度が 1 回、2 回の 2-gram の種類数である。

正規化定数$\lambda(\cdot)$は以下のように定義される。

$$\lambda(w_{i-1}) = \frac{d}{c(w_{i-1})} |\{w: c(w_{i-1},w) > 0\}|$$

ここで、$|\{w: c(w_{i-1},w) > 0\}|$は、$w_{i-1}$に後続する$w$の種類数である。

$P_{CONTINUATION}(w)$は以下で定義される。

$$P_{CONTINUATION}(w) = \frac{|\{w_{i-1}: c(w_{i-1},w) > 0\}|}{\left|\{(w_{j-1},w_j): c(w_{j-1},w_j) > 0\}\right|}$$

ここで、分母の$\left|\{(w_{j-1},w_j): c(w_{j-1},w_j) > 0\}\right|$は 2-gram の総種類数、分子の$|\{w_{i-1}: c(w_{i-1},w) > 0\}|$は、第 2 要素が$w$であるような 2-gram の種類数である。

文字列$w_1^l = w_1 w_2 \ldots w_l$と表される文書 q が年代tにおいて生成される真の尤度を$P_t(w_1^l)$、訓練データを用いた推定値を$\widehat{P_t}(w_1^l)$とすると、本モデルにおいて$\widehat{P_t}(q) = \widehat{P_t}(w_1^l)$は以下で与えられる。

$$P_t(q) \approx \widehat{P_t}(q) = \widehat{P_t}(w_1^l) = \prod_{i=1}^{l+1} \widehat{P_t}\left(w_i|w_0^{i-1}\right)$$

$$\approx \prod_{i=1}^{l+1} \widehat{P_t}(w_i|w_{i-1})$$

ここで、w_0は文頭記号、w_{l+1}は文末記号である。

年代tの事前確率に一様分布を想定すると、文書qの推定年代$\widehat{t_q}$は、対数尤度$\log \widehat{P_t}(w_1^l)$を最大にする年代$t$となる。

$$\widehat{t_q} = \arg\max_{t\in[1100,1700]} \log \widehat{P_t}(w_1^l)$$

$$= \arg\max_{t\in[1100,1700]} \log \prod_{i=1}^{l+1} \widehat{P_t}(w_i|w_{i-1})$$

$$= \arg\max_{t\in[1100,1700]} \sum_{i=1}^{l+1} \log \widehat{P_t}(w_i|w_{i-1})$$

6.2. *k*-近傍法（k Nearest Neighbors: kNN）

まず、各文書を$|F|$次元の二値ベクトルとして表現する。$|F|$は素性数である。素性セットは F1 と F2 をそれぞれ用いた。次に、年代推定を行う文書qと訓練データ内の各文書d_iとのコサイン類似度$\mathrm{Cos}(q, d_i)$を求める。全ベクトル要素の重みは同一とする。

$$\mathrm{Cos}(q, d_i) = \frac{\vec{q} \cdot \vec{d_i}}{|\vec{q}| \times |\vec{d_i}|}$$

文書qに対し最も高い類似度を示す k 個の文書の集合をS_kとし、文書qの推定年代$\widehat{t_q}$はS_kに属する文書$d' \in S_k$の実年代の加重平均とする（Manning *et al.* 2008: 273-275; Kawasaki 2014）。

$$\widehat{t_q} = \sum_{d' \in S_k} \left(t_{d'} \times \frac{\mathrm{Cos}(q, d')}{\sum_{d'} \mathrm{Cos}(q, d')} \right)$$

6.3. ナイーブベイズ分類器（Naive Bayes Classifier: NBC）

素性セットは F1 を用いる。年代推定は文書内での素性の有無のみに注目して行うので、多変数ベルヌーイモデルを採用した（高村 2010: 101-117）。文書 q が年代tにおいて生成される真の尤度を$P(q|t)$、訓練データを用いた推定値を$\hat{P}(q|t)$、$\hat{p}_{f,t}$を素性fの年代tにおける出現確率、$N_{f,t}$を素性fの出現する年代tの文書数、N_tを年代tの文書数、素性の集合をF、$\delta_{f,q}$を素性fが文書qに存在すれば 1、存在しなければ 0 を返す関数とする。

$\hat{p}_{f,t}$は MAP 推定で求め、平滑化パラメータ$\alpha_{f,t} \in \{1.001, 1.01, 1.1, 2.0\}$で実験を行った。

$$\hat{p}_{f,t} = \frac{N_{f,t} + (\alpha_{f,t} - 1)}{N_t + 2(\alpha_{f,t} - 1)}$$

年代tの事前確率$\hat{P}(t)$に一様分布を想定すると、事後確率$\hat{P}(t|q)$は、ベイズの定理と多変数ベルヌーイモデルを適用して、以下のように書き直すことができる。

$$\hat{P}(t|q) = \frac{\hat{P}(t)\hat{P}(q|t)}{\hat{P}(q)}$$

$$\propto \hat{P}(q|t) \approx \prod_{f \in F} \hat{p}_{f,t}^{\delta_{f,q}} \left(1 - \hat{p}_{f,t}\right)^{1 - \delta_{f,q}}$$

したがって文書qの推定年代$\widehat{t_q}$は、対数尤度$\log \hat{P}(q|t)$を最大にする年代tとなる。

$$\widehat{t_q} = \arg \max_{t \in [1100, 1700]} \log \hat{P}(q|t)$$

$$= \arg \max_{t \in [1100, 1700]} \log \prod_{f \in F} \hat{p}_{f,t}^{\delta_{f,q}} \left(1 - \hat{p}_{f,t}\right)^{1 - \delta_{f,q}}$$

$$= \arg \max_{t \in [1100, 1700]} \sum_{f \in F} \left(\delta_{f,q} \log \hat{p}_{f,t} + \left(1 - \delta_{f,q}\right) \log\left(1 - \hat{p}_{f,t}\right) \right)$$

7. 実験

7.1. 方法

訓練データ、テストデータともに、作成年代tが既知の文書のみから成る（全体で約 1400 文書）。テストデータに属する文書の作成年代tは不明だと仮定し、推定年代$\hat{t}$を求めた。N-gram モデルでは 15 分割交差検証を、*k*-近傍法（kNN）とナイーブベイズ分類器（NBC）では leave-one-out 交差検証（LOOCV）を用いた。実験はすべて Excel VBA で実装して行った。

7.2. 評価指標

評価指標は、絶対値誤差平均（Mean Absolute Error: MAE）、二乗平均平方根誤差（Root Mean Squared Error: RMSE）、絶対値誤差中央値（Median Absolute Error: MedAE）を用いる。離散変数である年代tを連続変数に読み替え、推定年代$\hat{t}$が実年代 t に近い（遠い）ほど良い（悪い）推定結果だとみなす。

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\widehat{t_i} - t_i| \; ; \; \mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\widehat{t_i} - t_i)^2}$$

ここで、Nはテストデータの文書数である。

7.3. 結果

表 1 に、各分類器の最良の結果を報告する。

表 1 実験結果

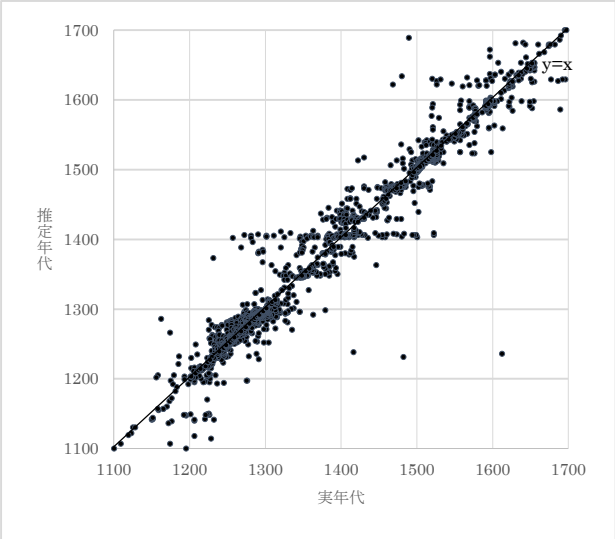
分類器	N-gram	kNN	kNN	NBC
素性セット	F2	F1	F2	F1
交差検証	15 分割	LOOCV	LOOCV	LOOCV
σ	10			5
k		6	6	
$\alpha_{f,t}$				1.01
MAE	**17**	21	20	25
RMSE	**31**	31	30	44
MedAE	**8**	14	13	14

MAE で評価した場合、N-gram モデルの成績が最も良く（約 17 年）、ナイーブベイズ分類器（NBC）が最も悪かった（約 25 年）。*k*-近傍法（kNN）では、用いた素性（F1 か F2）による MAE に統計的に有意な差はない（$p=0.549$）。いずれの分類器でも、パラメータ（$\sigma, k, \alpha_{f,t}$）の値は大きな影響を与えなかった。

8. 考察

図 3 は N-gram モデルでの実年代（横軸）と推定年代（縦軸）の散布図である。全体的には y=x の直線付近に分布しているが、特に 1400 年前後と 1620 年前後に系統的な誤推定が存在する。前者は言語的特徴が大きく異なる地方の文書の過度な集中、後者は文書長が極端に大きい文書の存在により引き起こされたものだと考えられる。このような系統的な誤推定は、*k*-近傍法では見られなかった。

図 3　N-gram モデルでの実年代（横軸）と推定年代（縦軸）の散布図



実は 1200 年以前の文書の大半はラテン語で書かれているため、これらの文書に中世スペイン語の素性セット（F1）を適用することはできない。また、言語システムが比較的安定している時代（17 世紀）においては、文献学的特徴のみで正確な年代推定を行うのは困難であり、文書内容の通時変化なども考慮する必要がある。一方、N-gram の素性セット（F2）は、言語に依存しないので複数の言語の文書集合にも適用でき、内容（たとえば語幹）の通時変化も捉えることが可能である。

年代推定の精度を高める方法としては、高次の N-gram モデルの構築、素性選択、誤差の大きい文書の除外、作成年代既知の文書と未知の文書を両方利用したナイーブベイズ分類器に対する半教師付き学習（Nigam *et al.* 2000）、文書数の増加等が考えられる。最後の点に関して、CODEA では年平均文書数が約 2 件と少ないうえ、文書長、発行地域、内容も不均質なため、パラメータの頑健な推定が困難となる。今後、コーパスの文書数は増加する予定であり、これにより推定精度の改善が期待される。

9. 結論

本論文では、N-gram モデル、*k*-近傍法、ナイーブベイズ分類器による中世スペイン語古文書の年代推定法を提案した。素性セットは文献学的特徴（F1）と文字 2-gram（F2）の二つの素性セットを用意した。実験の結果、N-gram モデルの成績が最も良く、絶対値誤差平均は約 17 年となった。

参考文献

Azofra, María Elena (2009): *Morfosintaxis histórica del español: de la teoría a la práctica*. Madrid: Universidad Nacional de Educación a Distancia.

Chen, Stanley, and Joshua Goodman (1998): An empirical study of smoothing techniques for language modeling. *Technical Report TR-10-98*, Harvard University.

CODEA = Sánchez-Prieto Borja, Pedro (dir.) (2010-): *Corpus de Documentos Españoles Anteriores a 1700*. http://demos.bitext.com/codea

Jurafsky, Dan and Christopher Manning. *Natural Language Processing*. https://class.coursera.org/nlp/lecture (as of 20/01/2015)

Kawasaki, Yoshifumi (2014): Datación crono-geográfica de documentos notariales medievales. *Scriptum Digital*. Vol. 3, pp. 29-63.

Manning, Christopher D., Prabhakar Raghavan and Hinrich Schütze (2008): *Introduction to information retrieval*. New York: Cambridge University Press.

Menéndez Pidal, Ramón (1999): *Manual de gramática histórica española* (vigésima tercera ed.). Madrid: Espasa-Calpe.

Nigam, Kamal, Andrew McCallum, Sebastian Thrun and Tom Mitchell (2000): Text classification from labled and unlabeled documents using EM. *Machine Learning, 39* (2/3), pp. 103-134.

Penny, Ralph (2002): *A history of the Spanish language* (Second Edition). Cambridge: Cambridge University Press.

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.

Tilahun, Gelila, Andrey Feuerverger and Michael Gervers (2012): Dating medieval English charters. *The Annals of Applied Statistics*, vol. 6 (4), pp. 1615-1640.