TYPO Wri✁ter: ヒトはどのように打ち間違えるのか？

荒牧英治† 宇野良子‡ 岡瑞起†
†東京大学
‡東京農工大学

eiji.aramaki@gmail.com ryokouno@cc.tuat.ac.jp mizuki.oka@gmail.com

1 はじめに

人間はどうして打ち間違えるのだろうか．優秀なタイピストであっても打ち間違いを起こす．また，何度も何度もチェックしたはずの原稿にタイポを発見してがっくりした経験はだれでも一度はあるだろう．いったい何が原因でタイポは生まれるのであろうか．

これまで，自然言語処理分野ではタイポは例外的現象として十分に研究されてこなかった．その理由は，新聞や教科書などこれまで言語処理で扱かっていた材料にそもそもタイポがほとんど存在しないことが大きい．しかし，膨大なテキスト資源であるウェブ上のサービス（ツィッターやチャットなど）は，出版物のような編集処理を経ておらず，その結果，非文法的表現やタイポを含んでいる．当然，これら新しいリソースを十分に利用するためには非文法的表現やタイポを処理できる頑健さが求められるであろう．

以上の背景から，我々は，タイポに焦点をあてて研究を行っている．その第一歩として，本研究では，大量のコーパスからタイポを自動抽出し（２章），それを分析することでタイポの原因を考察し（３章），定量的な検討を行う（４章）．

本研究の知見は，工学的にはタイポからその原型を推定する際の有用な手がかりとして利用可能であろう．また，タイポという人間の誤りを探ることによって，人間の認知メカニズムを探ることも期待できる．

2 タイポの抽出

まず，研究材料となるタイポを収集する．本研究では，タイポについて次の３つの仮定を設けて自動収集を行った．

仮定１「タイポの出現頻度はその原型の出現頻度に比べて著しく少ない」

仮定２「タイポは辞書に収載されていない*」

仮定３「タイポとその原型のスペリングは類似している」

2.1 材料

材料としてはツィッター（マイクロブログの一種；タイポが多く含まれいると見込まれる）†のクロールデータ約500MB‡を用いた．これらを，n-gram (n=1..3) に分解し，そこから小文字の英語だけを抽出した（大文字，数字／記号を含むものは，shift キーとの組み合わせやテンキーを用いての入力が考えられるため，対象外とした）．

表 1: 自動収集されたタイポの例.

原型	international	communication
タイポ	internationall	commmunication
	internatinal	communicaton
	inernational	comunication
	internationaly	communiction
	internaional	communiciation
原型	twitter	government
タイポ	teitter	gouvernment
	twitrer	governnment
	twixter	goverenment
	twittem	governement
	twittea	goverment

2.2 方法

まず，n-gram を集計し低頻度群（出現回数３回以下）と高頻度群（出現回数 100 回以上）を抽出した（仮定１による）．ここで，低頻度群をタイポ候補，高頻度群を原型候補と呼ぶことにする．次に，タイポ候補から辞書に収載されているものを除いた（仮定２による）．最後に，タイポ候補と原型候補のあらゆる組み合わせ（ペア）について，編集距離が１であるものを収集した（仮定３による）§．この際，低頻度群が複数の高頻度群の語と対応した場合は，もっとも出現頻度の高いものを原型として採用した．

2.3 結果

結果，38,200 組のタイポ:原型のペアを抽出することができた．例を表１に示す．抽出の精度であるが，正しいタイポというものが定義できないため，これを測定することは難しい．しかし，抽出された語は，実際に人間がタイプしたものであり，また，頻度が極めて少ないことから，語として流通していないことは明らかである．本研究では，これらをタイポとみなして考察をすすめる．

*この仮定により，本研究では，たまたまタイポが別の単語になってしまう場合（from→form など）を取り扱えない．よって，本研究はすべてのタイポ現象の一部を扱っていることとなる．

†http://twitter.com/

‡http://luululu.com/tweet/

§置換を編集距離１として扱った．

3　タイポの分析

いったいタイポはどのような原因で生成されているのであろうか．タイポは編集操作の観点からは次の 3 つに分類できる．

Insertion: 原型に余分な文字が加わる．

Deletion: 原型の文字が欠如する．

Replace: 原型の文字が別の文字に置換される．

表 2 に編集操作別のタイポが起こった文字を示す．表に示されるとおり，タイポが起きる割合は文字によって大きく異なり，決して均一に発生しているわけではないことが分かる．では，なぜこのような偏りが生じるのであろうか．

単語認知の関連研究をもとに以下の仮説を提案する（Hypothesis H1..H5).

【H1: 打鍵ミス】 打鍵する指がタイポと関連する．直接的には，タイポを生みだしているのはタイピングした指であり，運指との関連が考えられる．実際に，標準的な運指をした場合の各指のタイポ頻度（表 3）によれば，Replace が左手小指，薬指で多く発生していることが分かる．

【H2: 視覚的混同】 紛らわし文字がタイポを生む．画像として類似している文字がタイポの原因となっている可能性がある．例えば，文字画像の類似度 *sim*（図 1）では,「g」については「q」が,「m」については「n」がそれぞれ最も類似した文字となっており（表 4），実際にこれらのタイポ Replace「g→q」や「m→n」の頻度は高い（表 2）．

【H3: 単語内位置】 単語内の位置がタイポと関係する．英語において単語はスペースで区切られており，先頭または末尾にてタイポが発生した場合，非常に目につく．単語中でのタイポの位置を調査した結果を図 2 に示す．図にみられるように，単語先頭や末尾でのタイポ頻度は少なくなっている．

【H4: 文字冗長性】 周辺の文字との関連．表 1 上のような「communication」を「comunication」とタイポは極めて気がつきにくい．この理由として，原型には「m」が 2 つあり，そのうちの 1 つのみが deletion されているからである．このように，単語内に n 文字以上ある文字が欠落 (deletion された) した場合．または，語内にすでに存在した文字が Insertion される場合を本稿では文字冗長性を持つと呼ぶ．表 5 に示されるとおり，文字冗長性をもつタイポは多い．

【H5: 音韻的差異】 母音／子音の区別．Insert, Deletion ともに i,e,a など母音が多い．Replace においても e:a a:e,e:i といった母音間の交代が多い．このように，母音／子音の区別でタイポが説明できる可能性がある．

これらの仮説をまとめると，おおむね物理的原因 (H1), 視覚的原因（H2・H4），音韻的原因 (H5) などに大別することができる¶．もちろん，ある仮説ですべてのタイポが説明できるわけではなく，実際には，上記の仮説を中心にさまざまな原因でさまざまなタイポが生み出されているのが現実であろう．しかし，それぞれの仮説のおよその強さを探ることは可能であり，次章にて検討する．

¶これらの仮説はそれぞれ独立ではない．例えば,「単語先頭でのタイポは少ない (H2)」と「子音でのタイポは少ない (H5)」は単語先頭には子音が多いため独立でない．

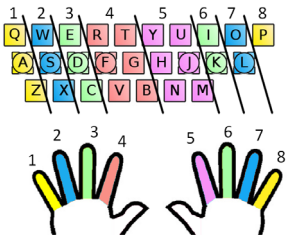
表 2: 編集操作別のタイポした文字.

Insertion		Deletion		Replace	
i	1858	g	0.08651 (=500/5790)	z:s	0.03061 (=32/1077)
e	1479	r	0.06746 (=890/13206)	x:s	0.02080 (=16/816)
a	1226	e	0.06428 (=1373/21371)	y:i	0.01695 (=78/4659)
r	910	d	0.06288 (=418/6662)	g:q	0.01689 (=140/8346)
n	856	y	0.05735 (=201/3521)	q:g	0.01631 (=6/428)
t	786	i	0.05321 (=884/16629)	c:k	0.01421 (=153/10831)
o	666	l	0.05294 (=479/9065)	a:e	0.01406 (=302/21538)
l	664	n	0.05206 (=766/14732)	m:n	0.01382 (=95/6945)
d	550	h	0.05142 (=301/5872)	d:r	0.01279 (=122/9614)
y	547	t	0.04979 (=661/13293)	k:c	0.01117 (=39/3579)
s	542	j	0.04878 (=23/491)	d:s	0.01109 (=106/9646)
h	453	c	0.04789 (=374/7828)	e:a	0.01087 (=331/30522)
u	409	p	0.04415 (=239/5435)	j:g	0.01057 (=6/661)
g	388	u	0.04395 (=230/5254)	x:n	0.01046 (=8/859)
c	353	o	0.04160 (=498/11993)	s:z	0.01045 (=179/17223)
z	342	a	0.04087 (=651/15949)	c:s	0.01027 (=110/10804)

* Deletion, Replace については割合（$\frac{Deletion\ または\ Replace\ された回数}{タイポコーパスにおけるその文字の出現数}$）でソートした（括弧内の値）

表 3: 編集操作別のタイポした指.

Insertion		Deletion		Replace	
5	2509	4	0.06143 (=1663/27083)	2:3	0.01982 (=425/21490)
3	2380	3	0.06036 (=2165/35883)	2:1	0.01546 (=332/21529)
6	2082	6	0.05156 (=978/18985)	1:3	0.01426 (=331/23275)
4	1815	5	0.04724 (=1555/32932)	2:5	0.01422 (=305/21507)
1	1662	7	0.04641 (=977/21071)	2:4	0.01100 (=236/21535)
7	1329	8	0.04415 (=239/5435)	6:3	0.01006 (=266/26526)
2	873	1	0.03964 (=676/17077)	8:5	0.00960 (=72/7596)
8	297	2	0.03539 (=546/15452)	6:5	0.00956 (=253/26557)



4　実験

本章では，各仮説を機械学習の素性としてインプリメントし，どの素性が貢献しているかを調査することで仮説の検証を行う．

4.1　実験設定

Support Vector Machine (SVM)[14] にて，タイポと非タイポを判別する学習器を構築する．素性には各仮説を定式化したものを用いる（表 6）．

正例：2 章にて抽出した原型：タイポのペア．

負例：正例の原型とそれをランダムに編集した（ランダムな位置でランダムに Insertion, Deletion, Replace を行ったもの）のペア．疑似負例．

実験にあたっては，7 文字のタイポ 5,000 ペアを正例とし，同数の負例を構築した．学習にあたっては，Tiny SVM‖にて多項式カーネル (d=2) を用いた**.

‖http://chasen.org/taku/software/TinySVM/

**t=1, d=2, c=1

表 4: 文字の視覚的類似度.

sim(g:*)				sim(m:*)			
高類似度		低類似度		高類似度		低類似度	
g:q	0.402	g:x	0.140	m:n	0.352	m:v	0.124
g:o	0.356	g:i	0.143	m:h	0.326	m:x	0.131
g:u	0.345	g:l	0.151	m:u	0.293	m:z	0.151
g:p	0.336	g:w	0.155	m:p	0.282	m:y	0.154
g:e	0.309	g:y	0.171	m:q	0.28	m:w	0.161
g:n	0.308	g:k	0.174	m:b	0.279	m:k	0.177
g:a	0.304	g:v	0.188	m:d	0.27	m:l	0.188
g:c	0.301	g:z	0.195	m:o	0.26	m:t	0.195
g:d	0.297	g:f	0.199	m:g	0.255	m:s	0.196

* 類似度は「MS ゴシック」フォントにて計算した.

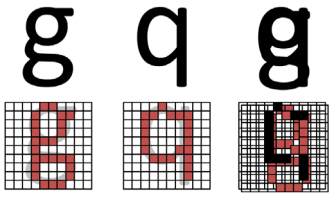


図 1: 文字の類似度. 文字「g」(左図) と「q」(中央) の面積とオーバーラップする部分の面積 (右図) の比で算出する.

$$sim(A,B) = \frac{\text{文字 } A \text{ と文字 } B \text{ のオーバーラップする面積}}{\text{文字 } A \text{ の面積} + \text{文字 } B \text{ の面積}}$$

ここで, 画像のオーバーラップを計算する際には, 最大 10%のずれを吸収し, オーバラップが最大になる位置で類似度を計算した.

比較手法は, すべての素性を使った場合を ALL とし, 検討したい素性を削除した手法 (with out H1〜5; w/o H1〜5) の精度を交差検定 (10 fold) にて比較した.

4.2 結果

結果, 表 7 を得た. どの素性を削除しても有意に精度が下がることから, 提案した仮説はいずれもタイポ現象の少なくとも一部を説明できると言える. また, 特にその中でも, H3,H4 の精度が低いことから, 単語内での位置による情報 (H3) と文字冗長性 (H4) がもっともよくタイポに影響していることが分かる.

この結果は次のように解釈できる. タイポを直接的に生み出しているのは打鍵する指であるが, その指の影響は比較的小さい. むしろ, タイポを抑制している視覚的要因の影響が大きく, 特に単語内での位置により抑制の精度は大きく変化する. また, 重複する文字などが存在する場合 (文字冗長性がある場合) などの影響も受ける.

ただし, 本研究から得られる知見には次の 3 つのバイアスの影響を受けている.

1. **タイポ自動抽出によるバイアス**: 自動抽出したタイポは, 辞書に収載されていないものだけであり (1 章仮定 2), 収集されたタイポに偏りが生じている可能性がある. しかし, 検証した仮説は辞書収載の有無と無関係であり, 大きなバイアスではないと考えている.

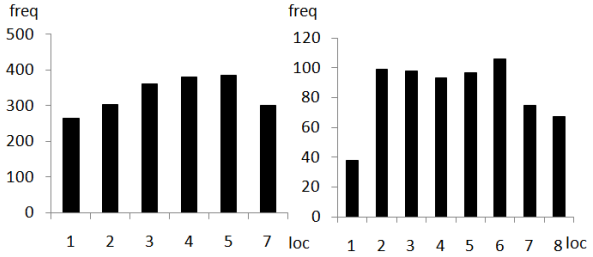


図 2: 7 文字の単語のタイポ位置 (loc 番目の文字) と頻度 (左). 8 文字の単語のタイポ位置と頻度 (右)

表 5: 文字冗長性を持つタイポの割合.

Deletion	0.3628 (=3479/9589)
Insertion	0.4760 (=6653/13976)

2. **仮説の定式化のバイアス**: 出現位置のように容易に素性に変換できる (定式化できる) 仮説もあれば, 運指のように定式化の方法が一意に定まらない仮説もある. このため, 大きな影響がないと判定された仮説も, 定式化で実験を行えば, 異なる結果が得られるかもしれない. しかし, 少なくとも本研究で強い影響を持つと示唆された仮説についてはその妥当性は損なわれないであろう.
3. **疑似負例によるバイアス**: 本研究で用いた負例は無作為に生成した疑似負例であり, この中に本来正例に分類されるべきものが混入している可能性がある. これにより, 不当に難しい分類課題を解くことになるが, すべての比較手法が同じ条件であるため, 大きなバイアスではないと考えている.

5 関連研究

これまでタイポの研究はされてこなかった. この理由は, そもそもこれまで扱った対象にタイポがほとんどないこともあげられる. しかし, OCR ／音声認識誤りや表記ゆれ吸収など関連の深い分野は存在する.

【認識誤りとの差異】近似文字列マッチング [7], スペル訂正 [3], 検索クエリ訂正 [4] . しかし, これらの認識誤りの対象は機械であり, 本研究のように人間の誤りを対象とはしておらず, スコープとは異なる.

【表記ゆれとの差異】表記ゆれを解消するために, 文字列同士の類似度を学習コーパスから獲得する研究が盛んに行われている. McCallum 等 [6] は, 編集操作を条件付き確率場で学習した. Tsuruoka ら [13] は, 文字列置換ルールの汎用性・曖昧性を計った. Aramaki 等 [2] や Okazaki 等 [10] は, 2 つの語の差分文字列を素性として, 表記揺れ識別器を構築した. 表記ゆれ現象は一定の規則性があり, それを機械学習することができる. 一方, タイポはその生成仮定の規則性は乏しく, 表記ゆれ技術を用いることはできない.

表 6: 素性と対応する仮説.

基本素性		仮説
基本素性	編集操作（Insertion, Deletion または Replace のうちのいずれか）とタイポされた文字.	-
タイポした指	指の番号 （表 3 による）	H1
文字画像類似度	Replace の場合，タイポした文字の類似度. Deletion または Insertion の場合，タイポした文字の面積.	H2
単語内位置	タイポが発生した位置.	H3
文字冗長性	文字冗長性を持つかどうかのフラグ.	H4
母音／子音	タイポした文字が母音か子音かのフラグ.	H5

* Replace の場合，B,H1,H3,H4,H5 は文字ペア（置換前の文字と置換後の文字）を扱った.

表 7: 結果.

	Accuracy	Precision	Recall
ALL	85.12%	78.13%	99.62%
w/oH1	82.20%	79.43%	85.51%
w/oH2	82.48%	80.35%	83.81%
w/oH3	73.90%	74.70%	73.70%
w/oH4	79.56%	76.19%	82.64%
w/oH5	82.40%	80.11%	83.74%

【単語認知との関連】タイポの抑制の観点からは単語認知の研究が関連する．文字の記憶は単語の先頭と末尾が強い（バスタブ効果）ことが知られている [1]．さらに，田中は情報量の観点からバスタブ効果を説明した [16]．本研究の知見は，これらの先行研究と整合するものである．

一方，音韻的観点からは，子音が語彙情報を担っていることが知られており [9]，実際に，子音に重点を置いて単語識別する傾向も報告されている [8]．しかし，タイポの知覚においては母音／子音の区別は顕著でなく，別のメカニズムが働いていることが示唆される．

【言語運用上の誤りとの関連】言語研究において，書き誤りや言い誤りは様々な問題意識から研究されている（Fromkin[5] を参照のこと）．特に単語レベルの書き誤りは発達や学習の観点から注目されてきた [11, 12]．言い直しのプロセスの場合は通常の発話データの一部に含まれるため，言い誤りの修正の観察データを比較的集めやすく，多くの研究がある（日本語の場合、例えば [15]）．一方で、タイピングで書くプロセスにおいては、書き誤りの修正はデータに通常残されないので研究対象とすることが難しい．本研究はこの困難を解決する手がかりとなる．

6 まとめ

本研究は，大量のコーパスからタイポを自動抽出し，タイポの原因を機械学習による素性として調査した．その結果，タイポが起こるためには，その位置や重複する文字があるかどうかという視覚的な要因が大きく関連することが明らかとなった．すなわち，タイポを生むのは指であるが，タイポを抑制するのは視覚であり，後者がタイポをタイポとして特徴づけている．今後，この知見をもとに，タイポを吸収するアルゴリズムの研究／開発や，人間の単語認知メカニズムの解明へと発展させていきたい．

参考文献

[1] J. Aitchison. *Words in the Mind*. Blackwell, 1994.

[2] Eiji Aramaki, Takeshi Imai, Kengo Miyo, and Kazuhiko Ohe. Orthographic disambiguation incorporating transliterated probability. In *Proceedings of International Joint Conference on Natural Language Processing (IJCNLP2008)*, pp. 48–55, 2008.

[3] E. Brill and R. C. Moore. An improved error model for noisy channel spelling correction. In *proceedings of the 38th Annual Meeting on Association for Computational Linguistics (ACL2000)*, pp. 286–293, 2000.

[4] Q. Chen, M. Li, and M. Zhou. Improving query spelling correction using web search results. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 181–189, 2007.

[5] V. A. Fromkin. *Errors in Linguistic Perfomance: Slips of the Tongue, Ear and Hand*. Academic Press Inc.

[6] A. McCallum, K. Bellare, and F. Pereira. A conditional random field for discriminatively-trained finite-state string edit distance. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 388–395, 2005.

[7] G Navarro. Acm computing surveys (csur). *A guided tour to approximate string matching*, Vol. 33, No. 1, 2001.

[8] T. Nazzi and B. New. Beyond stop consonants: Consonantal specificity in early lexical acquisition. *Cognitive Development*, Vol. 9, No. 22, 2007.

[9] M. Nespor, M. Pena, and J.Mehler. *On the different roles of vowels and consonants in speech processing and language acquisition, publisher =*.

[10] Naoaki Okazaki, Yoshimasa Tsuruoka, Sophia Ananiadou, and Jun'ichi Tsujii. A discriminative candidate generator for string transformations. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pp. 447–456, 2008.

[11] R Treiman. *Spelling*. Springer.

[12] R. Treiman, M. Cassar, and A. Zukowski. *What types of linguistic information do children use in spelling? The case of flaps*. 1994.

[13] Y. Tsuruoka, J. McNaught, and S. Ananiadou. Normalizing biomedical terms by minimizing ambiguity and variability. *BMC Bioinformatics*, Vol. 9, No. 3, 2008.

[14] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1999.

[15] 高梨克也, 丸山岳彦. 文と発話 3：時間の中の文と発話: 自発的な話し言葉に見られる挿入構造と線状化問題. ひつじ書房, 2007.

[16] 田中久美子. 単語に内在する情報量の偏在. 言語処理学会 第 14 回年次大会, pp. 1120–1123, 2008.