文字の出現確率に基づく第２類の文字列類似度

2006.2.26

情報コミュニケーション学会

飯箸泰宏

明治大学･法政大･慶応大講師

株式会社サイエンスハウス代表取締役

0. 問題意識

☆どの程度似ているのか？ –著作権上の紛争。

☆主観ではなく、再現性のある汎用的な尺度はないのか。

(サンプル)

- 「田子の浦ゆ うち出でて見れば 真白にそ　富士の高嶺に 雪は降りける」（万葉集）
- 「田子の浦に うち出でて見れば 白妙の 富士の高嶺に 雪は降りつつ」（新古今集）

1. 第1類と第2類の違い

- 第1類とは
- ランダムテキストに基づく文字列類似度

  ☆昨年、本学会で、発表し、「飯箸法」として一定の普及を見ることができた。

- 第2類とは
- 文字の出現頻度に基づく文字列類似度

  ☆本学会で、初めて公開する。

2.従来の測定法と飯箸法の違い

- **Levenshtein距離***

  2つの文字列を文字を削除、追加、入れ替えの操作を最少回数で同一にすることができる数

- **飯箸法****

  2つの文字列に共通するキーテキストがそれぞれの文字列にどの程度の重さを持っているかを計測する。

* A Method for the Correction of Garbled Words, Based on the Levenshtein Metric; K. Okuda, E. Tanaka, T. Kasai, IEEE Transactions on Computers, Vol. c-25, No. 2, February 1976

** 飯箸泰宏、”文字列類似度の汎用的尺度”、情報コミュニケーション学会第2回全国大会発表論文集、pp.53-54、2005年

3.文字列類似度の汎用的尺度

- テキスト間の相互類似度

$$rS_{a-b} = \frac{rS_{a/b} + rS_{b/a}}{2}$$

- テキスト間の相対類似度

$$rS_{a/b} = \frac{S_{a/b}}{O_a}, rS_{b/a} = \frac{S_{b/a}}{O_b}$$

- ⑴ 飯箸法第1類の類似度

$$\text{類似度} S_{a/b} = \ln(I_{a/b}) = \sum_{i=1}^{m_a} k_i \ln(N) - \ln\left(\prod_{i=1}^{m_a}(t_a - k_i + 1)\right)$$
$$= \sum_{i=1}^{m_a}(k_i \ln(N) - \ln(t_a - k_i + 1))$$
$$\text{類似度} S_{b/a} = \ln(I_{b/a}) = \sum_{i=1}^{m_b} k_i \ln(N) - \ln\left(\prod_{i=1}^{m_b}(t_b - k_i + 1)\right)$$
$$= \sum_{i=1}^{m_b}(k_i \ln(N) - \ln(t_b - k_i + 1))$$

$$\text{オリジナル度} O_a = \ln(L_a) = t_a \ln(N)$$
$$\text{オリジナル度} O_b = \ln(L_b) = t_b \ln(N)$$

- ⑵ 飯箸法第2類の類似度

$$\text{類似度} S_{a/b} = \ln(I_{a/b}) = -\sum_{i=1}^{m_a}\left(n_{1_{a_i}} \times \ln(p_{1_{a_i}})\right)$$
$$= -\sum_{i=1}^{m_a}\left(n_{1_{a_i}} \times \sum_{j=1}^{k_{a_i}} m_{c_{a_{ij}}} \times \ln(p_{c_{a_{ij}}})\right)$$
$$\text{類似度} S_{b/a} = \ln(I_{b/a}) = -\sum_{i=1}^{m_b}\left(n_{1_{b_i}} \times \ln(p_{1_{b_i}})\right)$$
$$= -\sum_{i=1}^{m_b}\left(n_{1_{b_i}} \times \sum_{j=1}^{k_{b_i}} m_{c_{b_{ij}}} \times \ln(p_{c_{b_{ij}}})\right)$$

$$\text{オリジナル度} O_a = -\ln\left(\prod_{i=1}^{t_a} p_i\right) = -\sum_{i=1}^{t_a} \ln(p_i)$$
$$\text{オリジナル度} O_b = -\ln\left(\prod_{i=1}^{t_b} p_i\right) = -\sum_{i=1}^{t_b} \ln(p_i)$$

- お詫び: 予稿集には、一部誤植がありました。このように訂正します。

4.尺度の比較

- 尺度の性格比較

	飯箸法 第2類	飯箸法 第1類	Levenshtein 距離
類似性	◎	○	×
距離	×	△	○
独立性	◎	○	△

飯箸法（第1類）とLevenshtein距離

この数は、独立性や類似性を意味しない。言葉を変えると「アモルファスな部分が増えると尺度としての意味がなくなる」のである

aaaaa!y_nと$z_n$aaaaa!（n=0,1,2,…,38）

n	文字列長	飯箸の尺度（第1類）		Levenshtein距離
		相互類似度 (%)	相互独立度 (%)	
0	6	100	0	0
1	7	84	16	2
	8	73	27	4
	9	64	36	6
4	10	57	43	8
5	11	54	46	10
6	12	47	53	12
7	13	43	57	13
8	14	40	60	14
9	15	37	63	15
…	…	…	…	…
38	44	12	88	44

独立性や類似性の意味がある数字である

5.実例

存在確率の小さな“s”が”st”という共通のキーテキストを作っているので、類似度は高くなる。

- 5-1 test とstreet

	飯箸法 第2類	飯箸法 第1類	Levenshtein 距離
相互類似性	86 %	25 %	
距離			3
相互独立性	14 %	75 %	

- 5-2

aaaaa! To be or not to be. That is question　と

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx aaaaa!

	飯箸法 第2類	飯箸法 第1類	Levenshtei n 距離
相互類似性	12 %	16 %	
距離			74
相互独立性	88 %	84 %	

存在確率が大きめの“a”が”aaaaa!”という共通のキーテキストを作っているので、類似度は低くなる。

- 5-3

「田子の浦ゆ うち出でて見れば 真白にそ　富士の高嶺に 雪は降りける」(万葉集)　と

「田子の浦に うち出でて見れば 白妙の 富士の高嶺に 雪は降りつつ」(新古今集)

	飯箸法 第2類	飯箸法 第1類	Levenshtein 距離
相互類似性	78 %	71 %	
距離			7
相互独立性	22 %	29 %	

存在確率が高い“に”などを除く確率の小さい文字によるキーテキストが多いので類似性は高くなる。

6.三法の比較考察

- 飯箸法の第1類は、文字の出現はランダムであると仮定した。
- 飯箸法の第2類は、文字の出現確率として比較空間の中で現に出現している文字の出現確率を用いる。
- 飯箸法の第1類では、文字の出現確率は文字種によって変化しないと見なしているので、この点では、Levenshtein法の「距離（文字数）」に近い考え方であった。
- 飯箸法の第2類では、文字の出現確率は文字種ごとに異なると見なしているので、Levenshtein法の「距離（文字数）」とはまったく類似性のない尺度となっている。

- 結果として、現実の文字列を見て感ずる人の感性的な類似性判断に飯箸法第2類は近づいたということができる。

7.まとめ

- 文字列の類似性など、従来人の感受性に頼ってきた分野に、「人の感性との間に矛盾がなく、かつ再現性のある」尺度を与えることが必要である。
- 再現性のある尺度という意味でLevenshtein距離には歴史的意味があった。
- Levenshtein距離に比べて、飯箸法第１類は、キーテキストの希少性の程度に着目することによって「類似性」という意味に沿った解を与えた。（「類似性」～「キーテキストの希少性の逆数の対数の和」）
- 飯箸法第１類に比べて、飯箸法第2類は、キーテキストの希少性を現実の文字種の存在確率に依拠することによってより現実感のある「類似性」の尺度となった。
- 今後は、よりいっそう人の感性との整合性のある尺度を探求する予定である。

終わり