E-016

Earth Mover's Distance を用いた同音異義語判別
Japanese Homophone Disambiguation Using Earth Mover's Distance

河原 直人† 梅澤 猛† 大澤 範高†

Naoto Kawahara† Takeshi Umezawa† Noritaka Osawa†

1. はじめに

情報機器への日本語入力において、入力誤りは大きく 2 種類に分ける事ができ、一つは文字単位で発生する入力誤りで、脱落、挿入、置換が挙げられる。もう一つは、単語単位で発生する仮名漢字変換における同音異義語誤りである。情報機器への日本語入力では仮名漢字変換が主流であるが、同じ読みを持つ変換候補から文脈に合ったものを選択する過程はユーザの目視に委ねられており、不注意や知識不足による誤りの発生がある。

そこで、本研究では同音異義語誤りを対象とし、同一文書内に出現する単語を分析することで文脈に則した単語を判別する手法を提案する。類似画像検索等で使用される距離尺度である Earth Mover's Distance(EMD)を用いることで、品詞情報と単語情報という、同音異義語判別に有用とされるが、本来異質であり同時に評価することが難しい情報を統一的に扱う事が可能になる。

2. Earth Mover's Distance

EMD は分布間の距離を表す尺度であり、類似画像検索などの分野で用いられている。EMD は分布間の距離の計算を輸送問題として捉え、最適な輸送コストを用いて定義される。EMD の算出には、まず需要地 P と供給地 Q をそれぞれ特徴量と重みのベクトルで表現する。次に特徴量間の輸送コスト $cost_{ij}$ を決定する。総輸送コストを最小化する輸送フローf_{ij} を決定することで、分布 P,Q 間の EMD を式(1)によって求めることができる。

$$\mathrm{EMD}(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} cost_{ij} f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} \quad (1)$$

需要地及び供給地の重みとその間の輸送コストを定義できれば、異質な特徴量が混在する場合や分布間の特徴量の数が異なる場合でも、EMD の値の算出が可能である。

竹内らは EMD を用いたキーワードによる画像検索を提案している[1]。従来より性能を向上させるために、キーワードに加えて画像自体から得られる特徴量を利用し、キーワード（単語）と画像特徴量という本来異質なものを扱う問題を、EMD の適用により解決している。柳本らはテキスト分類への EMD の応用を提案している[2]。EMD を用いることで索引語間の相関性を考慮したテキスト分類を可能としている。藤江らは概念ベースを用い単語間の関連性を求め、EMD により文書間の類似度を計算することによる文書検索への応用を提案している[3]。

3. 提案手法

EMD を同音異義語判別へ適応する手順について述べる。EMD を算出するためには、需要地と供給地の特徴量とその重み、輸送コストを定義する必要がある。

† 千葉大学大学院融合科学研究科 Graduate School of Advanced Integration Science, Chiba University

3.1 需要地・供給地の定義

まず、需要地と供給地を定義する。需要地には判別対象を含む文書、供給地には仮名文字列 h に対する変換候補 $h_1,h_2,\cdots h_n$ それぞれについて特徴量と重みを設定する。

同音異義語判別のための特徴量として、局所的情報と大域的情報を割り当てる。局所的情報には対象とする同音異義語の直前と直後に出現する単語の品詞情報の組み合わせを利用する。h_k に対する供給地の局所的情報には、同音異義語 h_kが出現する文書群 D_kから同音異義語 $h_1,h_2,\cdots h_n$に関する品詞情報を抽出することになる。例えば次のような文があったとき、同音異義語『内蔵』に関する品詞情報は『名詞 - 助詞』となる。

（例）タグ **内蔵** の 3D 加速度センサーが振動を感知すると
名詞 助詞

大域的情報には文書もしくは文書群に出現する名詞、動詞、形容詞の単語を特徴量として用いる。

特徴量の重みは文書 d_j における単語 t_i（または品詞情報 s_i）の重み$w(t_i,d_j)$を *tf-idf* より以下の式(2)で求める。

$$w(t_i,d_j) = f(t_i,d_j)\ \log\frac{N}{g(t_i,d_j)} \quad (2)$$

ここで、$f(t_i,d_j)$は d_jにおける単語 t_iの出現頻度、Nは全文書数、$g(t_i,d_j)$は単語 t_i を含む文書数を表す。D_k における単語 t_j の重み$w(t_i,D_k)$は、$f(t_i,D_k)$を D_k における単語 t_i の出現頻度、$g(t_i,D_k)$は t_iを含む文書数とし定義する。

3.2 輸送コストの定義

輸送コストとして局所的情報 - 局所的情報間、大域的情報 - 大域的情報間、局所的情報 - 大域的情報間（大域的情報 - 局所的情報間）の 3 種のコストを定義する必要がある。コストは 0 から 1 の値になるよう決定する。

局所的情報と局所的情報間のコストは品詞情報が合致するかを基に、式(3)で定義する。

$$cost(s_i,s_j) = \begin{cases} 0 & (s_i = s_j) \\ 1 & (s_i \neq s_j) \end{cases} \quad (3)$$

次に大域的情報と大域的情報間のコストを単語の出現文書数に基づいて以下の式(4)で定義する。

$$cost(t_i,t_j) = 1 - \frac{n_{ij}}{\sqrt{n_i n_j}} \quad (4)$$

ここで、n_{ij} は単語 t_i と t_j の共起文書数、n_i, n_j は t_i, t_j の出現文書数である。式(4)を用いて局所的情報と大域的情報間のコストも出現文書数から求める。

3.3 EMD による同音異義語判別

先に定義した需要地、供給地の重みと輸送コストを用いることで、各変換候補文字列に対する EMD を式(1)から算出し、EMD の最も小さな変換候補を判別結果とする。

4. 実験

提案手法の有用性を検証するために、『内蔵／内臓』および『意外／以外』の 2 組の同音異義語について判別実験を行った。

4.1 供給地と輸送コスト

収集した 15 文書（以下文書群 D と呼ぶ）を基に、供給地の特徴量と重み及び輸送コストを算出する。大域的情報として利用した単語（名詞・動詞・形容詞）は 1282 語であり、局所的情報として利用した品詞情報は、同音異義語『内蔵／内臓』における組み合わせが 3 種類、同音異義語『意外／以外』における組み合わせが 4 種類であった。D に出現した品詞情報を表 1 に示す。

表 1　文書群 D に含まれる品詞情報

内蔵／内臓			意外／以外		
品詞情報	出現文書数	出現頻度	品詞情報	出現文書数	出現頻度
名詞-助詞	1	1	記号-名詞	1	1
記号-名詞	5	6	助詞-助詞	1	1
助詞-名詞	1	1	記号-助動詞	1	1
			名詞-助詞	4	5

仮名文字列『ないぞう』は 1285 個、『いがい』は 1286 個の特徴量と重みを供給地として持つことになる。輸送コストは全ての特徴量の組み合わせについて(3), (4)の式を用いて算出する。

4.2 需要地

需要地には判別対象とする文字列を含む文書を D とは別に用意した。判別対象文書に出現する単語（名詞・動詞・形容詞）と対象文字列に対する品詞情報を特徴量とし重みを算出する。ここで、D には出現しない単語や品詞情報が特徴量として抽出される可能性がある。この時、新出の特徴量については輸送コストが定義されていないため、それらに関する全ての輸送コストを 1 と定義した。

特徴量の数は需要地の方が多くなるが、EMD は特徴量の数が異なる場合にでも求めることができるので問題にはならない。変換候補それぞれに対する EMD を求め、その文書において正解となる漢字文字列を決定する。

5. 結果と考察

『内蔵』または『内臓』が正解となる文書群を A、『意外』または『以外』が正解となる文書群を B とする。D、A、B について文字数及び単語数の平均値、最大値、最小値を表 2 に示す。

表 2　各文書群の文字数及び単語数

	D		A		B	
	文字数	単語数	文字数	単語数	文字数	単語数
平均値	638.8	399.1333	829.4	503.2	978.5	623.6
最大値	957	602	2281	1263	2261	1451
最小値	157	100	254	171	467	303

5.1 『内蔵』と『内臓』の判別

『内蔵』が正解の 5 文書 $a_1,a_2,\cdots,a_5$ の結果を表 3、『内臓』が正解の 5 文書 $a_6,a_7,\cdots,a_{10}$ の結果を表 4 に示す。

表 3 『内蔵』を含む文書の EMD 値

	a_1	a_2	a_3	a_4	a_5
EMD(内蔵)	**0.5123**	**0.61304**	**0.66227**	**0.78461**	**0.73775**
EMD(内臓)	0.892645	0.87591	0.87616	0.823135	0.839418

表 4 『内臓』を含む文書の EMD 値

	a_6	a_7	a_8	a_9	a_{10}
EMD(内蔵)	0.862648	0.853993	0.859193	0.861999	0.889918
EMD(内臓)	**0.66469**	**0.59367**	**0.44065**	**0.77528**	**0.79553**

5.2 『意外』と『以外』の判別

『意外』が正解の 5 文書 $b_1,b_2,\cdots,b_5$ の結果を表 5、『以外』が正解の 5 文書 $b_6,b_7,\cdots,b_{10}$ の結果を表 6 に示す。

表 5 『意外』を含む文書の EMD 値

	b_1	b_2	b_3	b_4	b_5
EMD(意外)	**0.82946**	**0.82383**	**0.88989**	0.845995	0.880878
EMD(以外)	0.836736	0.8285	0.894553	**0.83648**	**0.87938**

表 6 『以外』を含む文書の EMD 値

	b_6	b_7	b_8	b_9	b_{10}
EMD(意外)	0.851516	0.884121	**0.85109**	**0.85909**	0.88688
EMD(以外)	**0.82808**	**0.80775**	0.85134	0.861717	**0.86633**

それぞれの正判別率を表 7 にまとめる。

表 7 正判別率

	内蔵	内臓	意外	以外	全体
正解数	5	5	3	3	16
正判別率(%)	100	100	60	60	80

5.3 考察

『内蔵／内臓』の判別は 100%という高い正判別率を得ることができた。これは『内蔵』が含まれる文書にはコンピュータ関連の内容が、『内臓』が含まれる文書には健康に関する内容が書かれているなど、共起する単語に特徴が現れやすいことが原因と考えられる。EMD の値も比較的差のある結果を得ることができた。

一方、『意外／以外』の正判別率は『内蔵／内臓』に比べると低い。このような大域的情報（単語）による EMD の差が生じない場合は、局所的情報（品詞情報）へ依存した判別をすることが望ましい。そこで、大域的情報に比べ局所的情報の重みを大きく設定するといった改善が有効ではないかと考えられる。例えば文書 b_5 は判別に失敗しているが品詞情報は『記号 - 助動詞』であり、正解である『意外』の特徴量とのみ一致していた。この時、局所的情報に大きな重みを与えれば『意外』に関する EMD は小さくなることが見込め、正判別率の向上が期待できる。

6. おわりに

本稿では EMD を用いた同音異義語判別について特徴量やその重み、輸送コストの定義について提案した。実験の結果、正判別率 80%を得ることができ、EMD を用いた同音異義語判別が有用であることの示唆が得られた。今後は局所的情報の重みを大きくするような新たな重み付けの定義や、文書数を増やした実験へ取り組む予定である。

参考文献

[1] 竹内 謹治, 黄瀬 浩一, “Earth Mover's Distance に基づく Text-Based Image Retrieval”, 情報処理学会研究報告. NL-1775-5, pp.33-38, (2007).

[2] 柳本 豪一, 大松 繁, “Earth Mover’s Distance を用いたテキスト分類”, 人工知能学会全国大会論文集(CD-ROM), 21st, pp.1G3-4, (2007).

[3] 藤江 悠五, 渡部 広一, 河岡 司, “概念ベースと Earth Mover's Distance を用いた文書検索”, 情報処理学会研究報告. ICS, 2009(16), pp.111-116, (2009).