# 単語ベクトルの長さは意味の強さを表す

大山 百々勢 *,1,3　横井 祥 *,2,3　下平 英寿 *,1,3
[1] 京都大学　[2] 東北大学　[3] 理化学研究所
oyama.momose@sys.i.kyoto-u.ac.jp　yokoi@tohoku.ac.jp
shimo@i.kyoto-u.ac.jp

## 概要

単語ベクトルは単語の持つ様々な言語的特性を反映していると考えられるが，単語ベクトルの長さに何がどのようにエンコードされているかは明らかでない．本稿では，単語ベクトルの長さが意味の強さをエンコードするという経験的な知見に理論的な枠組みを与える．はじめに意味の強さの指標として，各単語の周辺単語分布とコーパス全体の単語分布の Kullback-Leibler (KL) 情報量を提案する．またこの指標の妥当性を，単語頻度によるバイアスを KL 情報量から適切に除去した尺度が品詞の違いを識別できるという実験を通して確かめる．さらに，Skip-gram with Negative Sampling によって得られる単語ベクトルの長さの 2 乗が KL 情報量，つまり意味の強さに相当することを理論と実験で示す．

## 1　はじめに

自然言語処理のタスクにおいて欠かせない単語ベクトルには，単語の持つ各種意味的・統語的特性がエンコードされていることがわかっている [1, 2, 3]．本稿ではとくに単語ベクトルの長さ（ユークリッドノルム）に関して理論的・経験的分析をおこなう．

単語ベクトルの長さは「単語の意味の強さ」を反映すると既存研究でも考えられている．単語ベクトルの加法構成性を拡張して意味に関する論理演算（AND，OR，NOT）を明らかにした先行研究 [4] によれば，単語ベクトルの加算は AND 演算に相当する．したがって，ある単語を複数回足し合わせてベクトルを長く することが意味を強める操作に対応すると予想される．また [5] によれば，固有名詞の方がより「弱い意味」を持つであろう機能語よりも単語ベクトルが長いという実験結果が示されている．

本稿では「意味の強さ」という曖昧な対象を数理的に記述するため，その指標として**注目している単語の周辺単語分布とコーパス全体の単語分布の Kullback-Leibler (KL) 情報量**を採用する．KL 情報量はその単語が出現したという情報の大きさを表しており，「意味の強さ」の指標として自然な選択肢だと考えられる．この定式化によって，単語ベクトルの長さが意味の強さを表すという仮説に対して様々な実証実験や理論保証が可能となる．たとえば，単語ベクトルの長さは KL 情報量と実際に密接な関係があることが実験的に見て取れる（図 1）．また後ほど 3 節で示すように，ある特殊な白色化処理を施した単語ベクトルの長さの 2 乗は，KL 情報量（つまり意味の強さ）の近似値であること示せる（図 4）．

経験的な検証にあたっては，単語頻度（コーパスにおける単語の出現回数）の影響が状況を複雑にする．頻度と KL 情報量には強い相関があり，頻度が大きい単語ほど KL 情報量は小さい傾向がある（図 1）．より多く使われる単語のほうが特殊性が低く意味が弱い傾向があると考えればごく自然なことである．さらに，KL 情報量は有限長のコーパスから計算するため，とくに低頻度語では量子化誤差やサンプリング誤差の強い影響をうける．そこで本論文では，KL 情報量から頻度の直接的影響を除去した頻度補正済み KL 情報量を尺度として用いる．結果，固有名詞のほうが機能語や動詞より頻度補正済み KL 情報量の値が大きい傾向があることを，つまり意味が強い傾向があることをより精確に確認することができた．すなわち，KL 情報量を用いた「長さ＝意味の強さ」説の理論保証は，頻度のバイアスによる擬似相関でないことが確認できた．

## 2　単語の KL 情報量と頻度の影響

### 2.1　単語の KL 情報量

コーパスにおける単語 $w$ の確率を $p(w)$ とし，コーパス分布を $p(\cdot)$ と表す．同様に，単語 $w$ の周辺（または文脈）単語が $w'$ である確率を $p(w'|w)$ とし，

∗ Contributed equally

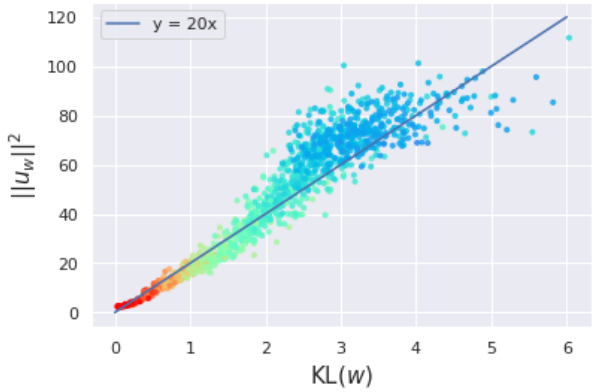


**図 1**　KL 情報量と単語ベクトルの長さの 2 乗の散布図．各点は text8 コーパスから選んだ 1200 単語であり，単語頻度を色で表す．高頻度語は暖色，低頻度語は寒色．詳細は 2.3 節，3.2 節を参照．

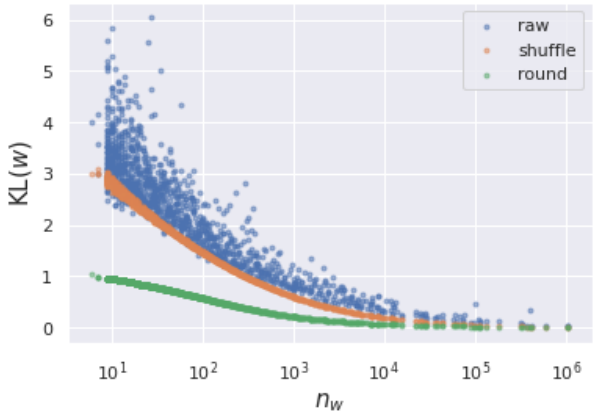


**図 2**　単語頻度 $n_w$ と 3 種類の KL 情報量の散布図: KL($w$) (raw), $\overline{\mathrm{KL}}(w)$ (shuffle), $\mathrm{KL}_0(w)$ (round)．各点は text8 コーパスから選んだ 1200 単語．

周辺単語分布を $p(\cdot|w)$ と表す．分布仮説 [6, 7] によれば，周辺単語分布 $p(\cdot|w)$ は単語 $w$ の意味を表すとされる．これらは確率分布であるから，語彙集合を $V$ とすれば，$\sum_{w\in V} p(w) = \sum_{w'\in V} p(w'|w) = 1$ である．$p(\cdot|w)$ の重み付き平均（または重心）がコーパス分布 $p(\cdot)$ であり，$p(\cdot) = \sum_{w\in V} p(\cdot|w)p(w)$ である．

2 つの分布 $p(\cdot|w)$ と $p(\cdot)$ の KL 情報量（または KL ダイバージェンス）は

$$\mathrm{KL}(p(\cdot|w) \parallel p(\cdot)) = \sum_{w'\in V} p(w'|w) \log \frac{p(w'|w)}{p(w')}$$

と定義される．本論文では，これを KL($w$) と表し，単語 $w$ の KL 情報量とよぶ．KL($w$) は $w$ の周辺単語分布がいかに重心からズレて特殊なものになっていくかを表していて，単語 $w$ のもつ情報の大きさと考えられる．

## 2.2　単語頻度の直接的影響

KL 情報量は単語頻度と高い相関がある．実験設定は 2.3 節で説明するが，図 2 の raw は，単語 $w$ の出現回数 $n_w$（これを単語頻度とよぶ）と KL($w$) の関係を表している．単語頻度が小さいほど KL 情報量が大きくなる傾向がわかる．この傾向の大きな要因は，コーパスが有限長 $N$ であることによって生じる誤差であり，一種のアーティファクトまたはバイアスといえる．この誤差を本論文では，単語頻度の直接的影響とよぶ．このような影響を適切に取り除いた尺度を用いることで，正しく単語の意味の強さが測定できる．

単語分布 $p(\cdot)$ と $p(\cdot|w)$ は有限長のコーパスから計算される．単語 $w$ の確率は $p(w) = n_w/N$ である．単語 $w$ の前後 $\pm h$ の窓における単語 $w'$ の出現回数，すなわち共起回数を $n_{w,w'}$ とすると，周辺単語 $w'$ の確率は $p(w'|w) = n_{w,w'}/\sum_{w''\in V} n_{w,w''}$ で計算される（ただし，コーパスの端点を無視すれば，$\sum_{w''\in V} n_{w,w''} = 2hn_w$）．

### 2.2.1　サンプリング誤差

もとのコーパスをランダムにシャッフル，すなわち単語を並べ替えたコーパスを考える [8]．$n_w$ とそれから計算される $p(\cdot)$ は不変であるが，単語の共起回数 $n_{w,w'}$ とそれから計算される $p(\cdot|w)$ は単語の意味を反映しなくなる．このようにして作った $n_{w,w'}$ は，コーパス端点の影響を無視すれば，$p(\cdot)$ をパラメータにもつ多項分布からのサンプリングに等価である．このランダムなコーパスから計算した単語 $w$ の KL 情報量の平均値を $\overline{\mathrm{KL}}(w)$ とする．実験では，ランダムにシャッフルしたコーパスを 10 個作成し，その KL($w$) の平均値を $\overline{\mathrm{KL}}(w)$ として用いる．これは単語の意味は反映せず，サンプリングによるランダムネスの影響のみを表している．図 2 の shuffle に示すように，$\overline{\mathrm{KL}}(w)$ は $n_w$ が小さいほど大きくなる．KL 情報量から頻度の直接的影響を取り除いた $\mathrm{KL}(w) - \overline{\mathrm{KL}}(w)$ を頻度補正済み KL 情報量とよぶことにする．本論文では，これを意味の強さの尺度として用いる．

### 2.2.2　量子化誤差

低頻度語 $w$ では $n_w$ が小さいため，$n_{w,w'} = 0$ となる $w'$ が増えてスパースになり，回数が整数値しかとらないことによる量子化誤差の影響も無視

できなくなる．そこで $n_{w,w'} := \mathrm{round}(n_w p(w')2h)$ と再定義して計算した KL 情報量を $\mathrm{KL}_0(w)$ とし，図 2 の round に示す．整数値の丸め誤差がなければ，$p(w'|w) = p(w')$ となり $\mathrm{KL}_0(w) = 0$ のハズである．しかし実際には，$n_w$ が小さい単語では $\mathrm{KL}_0(w)$ は無視できない大きさになっている．3 節で単語ベクトルの長さと KL 情報量の関係を確認するとき，$\mathrm{KL}(w) - \mathrm{KL}_0(w)$ を使うことで，KL 情報量のゼロ点調整を行う．

### 2.3 実験

KL 情報量から単語頻度の直接的影響を除去した尺度を計算し，単語の意味の強さがこの尺度の大小に反映されているか確認する．text8 コーパスの単語共起行列から KL 情報量を計算する．単語集合 1 はコーパスからバランス良く選んだ 1200 単語，単語集合 2 は固有名詞 10561 単語，機能語 123 単語，動詞 4771 単語からなる．実験設定の詳細は付録 B を参照．

**■実験 1** 単語集合 1 の各単語 $w$ について $\mathrm{KL}(w)$ を計算する．SGNS[9] で計算した 300 次元の単語ベクトル $u_w$ の長さの 2 乗 $\|u_w\|^2$ との関係を図 1 に示した．単語頻度の直接的影響を図 2 に示した．単語ベクトルの詳細な結果は 3.2 節で示す．

**■実験 2** 単語集合 2 の各単語 $w$ について $\mathrm{KL}(w) - \overline{\mathrm{KL}}(w)$ を計算した結果を図 3 に示す．固有名詞は，機能語や動詞よりも大きな値になる傾向がある．固有名詞の文脈は機能語や動詞の文脈に比べて限定的と考えられ，相対的に意味が強くなりやすいと期待される [5] ことと矛盾無い結果が得られた．固有名詞のうち出現回数が 100 回以上 1000 回未満の 2428 単語について，$\mathrm{KL}(w) - \overline{\mathrm{KL}}(w)$ の値でソートした上位，中位，下位の各 10% 区間の単語から，それぞれ 10 単語をランダムにサンプリングしたものを表 1 に示す．この処理では大文字小文字の区別はしておらず，storm, haven などむしろ普通名詞とすべきものが含まれている．このような普通名詞が下位 240 単語には多く見られる．一方，上位 240 単語には企業名などの固有名詞が多く見られる．

## 3 KL 情報量と単語ベクトルの長さ

図 1 で見たように，単語ベクトルの長さの 2 乗は KL 情報量の近似であることが予想される．このことを理論的，実験的に示す．

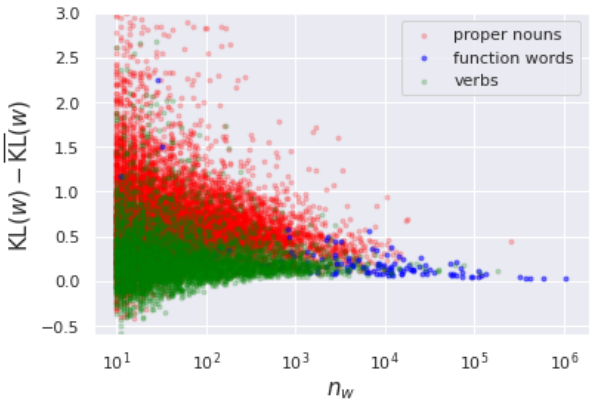


**図 3** 単語頻度 $n_w$ と意味の強さの尺度 $\mathrm{KL}(w) - \overline{\mathrm{KL}}(w)$ をプロットした．固有名詞は 10561 単語（赤），機能語は 123 単語（青），動詞は 4771 単語（緑）である．

| $\mathrm{KL}(w) - \overline{\mathrm{KL}}(w)$ | 単語例 |
|---|---|
| 上位 (0% ~ 10%) | HONDA, INTERPOL, Gabon, Yin, VAR, IMF, Benin, BO, Bene, GB |
| 中位 (45% ~ 55%) | Pete, Dee, Wine, Tony, Bogart, Alice, Cliff, Madonna, Dover, Leopold |
| 下位 (90% ~ 100%) | storm, haven, sale, miracle, discover, Phillip, duty, prohibition, capitol, comfort |

**表 1** text8 コーパス内での出現回数が 100 回以上 1000 回未満の名詞・固有名詞について，$\mathrm{KL}(w) - \overline{\mathrm{KL}}(w)$ の値の上位・中位・下位それぞれ 240 単語からランダムに 10 個ずつサンプリングした．

### 3.1 理論

SGNS では 2 種類の単語ベクトルを学習する．単語側埋め込みを $u_w, w \in V$，文脈側埋め込みを $v_{w'}, w' \in V$ とする．これらの重心を $\bar{u} = \sum_{w \in V} p(w) u_w$, $\bar{v} = \sum_{w' \in V} p(w') v_{w'}$ とし，頻度重み付き中心化した単語ベクトルを $\hat{u}_w := u_w - \bar{u}_w, \hat{v}_{w'} := v_{w'} - \bar{v}$ とする．

SGNS の negative sampling 分布を $q(\cdot)$，1 共起当たりの負例数を $\nu$ とする．理想的な学習で単語ベクトルが得られていると仮定すれば，周辺単語分布は $p(w'|w) = \nu q(w') e^{\langle u_w, v_{w'} \rangle}$ である．議論を簡単にするため $p(\cdot) = q(\cdot)$ と仮定すると，KL 情報量は

$$\mathrm{KL}(w) \simeq \frac{1}{2} u_w^\top \mathrm{Cov}(v) u_w$$

ただし $\mathrm{Cov}(v)$ は文脈側埋め込みの頻度重み付き分散共分散行列である．この導出は付録 A を参照．$u = 0$ が $q(\cdot)$ に，$u = \bar{u}$ が $p(\cdot)$ に対応することを考慮すると，一般に $p(\cdot) \neq q(\cdot)$ のときは $u_w$ を $\hat{u}_w$ でおき

かえたほうがよい．そこで，$\hat{u}_w$ を**文脈側ベクトルで白色化**[1]したベクトルを $\tilde{u}_w := \mathrm{Cov}(v)^{\frac{1}{2}}\hat{u}_w$ とおけば，$\|\tilde{u}_w\|^2 \simeq 2\mathrm{KL}(w)$．

### 3.2 実験

text8 コーパスから SGNS を用いて 300 次元の単語ベクトル $u_w, v_w$, $w \in V$ を学習した．反復数は 100 エポックとした．量子化誤差の影響（2 節を参照）を考慮するためにゼロ点調整を行った KL 情報量を用いる．単語集合 1 の 1200 単語について KL 情報量と $\|\tilde{u}_w\|^2$ の関係を図 4 に示す．$\|\tilde{u}_w\|^2 \simeq 2(\mathrm{KL}(w) - \mathrm{KL}_0(w))$ の関係が比例定数を含めて近似的に成立していることが確認できる．

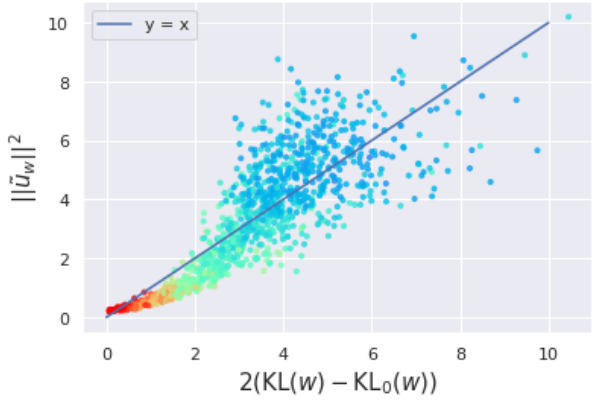


**図 4** 縦軸は文脈ベクトルで白色化した単語ベクトル $\tilde{u}_w$ の長さの 2 乗．横軸は $2(\mathrm{KL}(w) - \mathrm{KL}_0(w))$．その他の設定は図 1 と同じ．

## 4 単語ベクトルの長さと意味の強さ

2 節と 3 節の結果から，単語ベクトルの長さに単語の意味の強さがエンコードされていることが示唆される．これを実験的に確認する．KL 情報量と同様に単語ベクトルの長さにもサンプリング誤差があるため，2.2.1 節の方法で単語頻度の直接的影響を取り除く．ランダムにシャッフルしたコーパスから学習した $\|u_w\|^2$ の平均値を $\overline{\|u_w\|^2}$ とする．ここでは 10 回の平均値を用いる．

図 5 は単語ベクトルの頻度補正済み長さの 2 乗 $\|u_w\|^2 - \overline{\|u_w\|^2}$ を図 3 と同様にプロットしたものである．KL 情報量を用いたときとまったく同様に，機能語，動詞に比べて固有名詞の意味が強い傾向が読み取れる．$u_w$ の代わりに文脈ベクトルで白色化した単語ベクトル $\tilde{u}_w$ を用いた場合を図 6 に示す．$\tilde{u}_w$ を用いると KL 情報量 ×2 に相当するため解釈が容易になる反面，低頻度領域で若干のバイアスが残っているように見える点は今後の課題である．

1） 本稿では便宜上「白色化」と呼んでいるが，変換式において $\mathrm{Cov}(v)$ の次数は $-\frac{1}{2}$ ではなく $\frac{1}{2}$ であることに注意されたい．

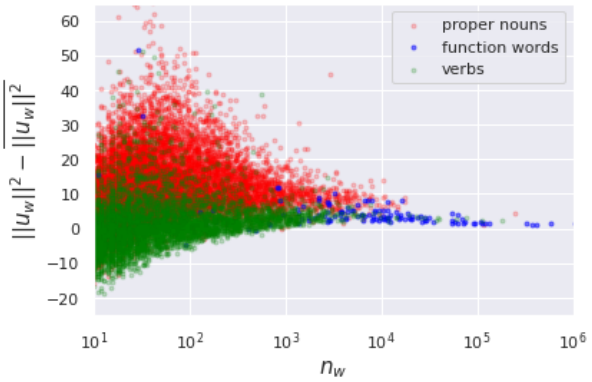


**図 5** 単語頻度 $n_w$ と単語ベクトルの頻度補正済み長さの 2 乗 $\|u_w\|^2 - \overline{\|u_w\|^2}$ のプロット．

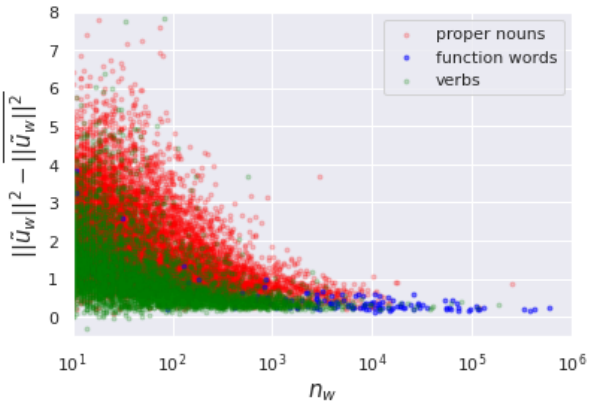


**図 6** 単語頻度 $n_w$ と文脈ベクトルで白色化した単語ベクトルの頻度補正済み長さの 2 乗 $\|\tilde{u}_w\|^2 - \overline{\|\tilde{u}_w\|^2}$ のプロット．

## 5 おわりに

単語の KL 情報量を「意味の強さ」と解釈することを提案した．単語の品詞の違いや，名詞・固有名詞の違いが矛盾なく意味の強さに反映されていることを数値例で確かめた．単語の KL 情報量の代わりに単語ベクトルの長さの 2 乗を用いても同じように意味の強さが測れることを確かめた．これらの評価では単語頻度の影響を適切に除去することが重要であり，そのためにランダムにシャッフルしたコーパスを用いて KL 情報量を補正した．今後は，WordNet[10] 等のデータベースを用いて単語の上位・下位関係，含意関係などがどのように反映されているか確かめたい．

## 謝辞

本研究は，JSPS 科研費 20H04148, JST CREST JPMJCR21N3，JST ACT-X JPMJAX200S の助成を受けています.

## 参考文献

[1] Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. Evaluation methods for unsupervised word embeddings. In **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 298–307, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[2] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 842–866, 2020.

[3] Sho Yokoi, Ryo Takahashi, Reina Akama, Jun Suzuki, and Kentaro Inui. Word rotator's distance. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2944–2960, Online, November 2020. Association for Computational Linguistics.

[4] Masahiro Naito, Sho Yokoi, Geewook Kim, and Hidetoshi Shimodaira. Revisiting additive compositionality: AND, OR and NOT operations with word embeddings, 2021. ACL-IJCNLP 2021 Student Research Workshop.

[5] Adriaan M. J. Schakel and Benjamin J. Wilson. Measuring word significance using distributed representations of words, 2015. arXiv 1508.02297.

[6] Zellig Harris. Distributional structure. **Word**, Vol. 10, No. 2-3, pp. 146–162, 1954.

[7] J. R. Firth. A synopsis of linguistic theory 1930-55. **Studies in Linguistic Analysis (special volume of the Philological Society)**, Vol. 1952-59, pp. 1–32, 1957.

[8] Kumiko Tanaka-Ishii. **Statistical Universals of Language**. Springer, 2021.

[9] Tomás Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In **Advances in Neural Information Processing Systems**, pp. 3111–3119, 2013.

[10] Christiane Fellbaum. **WordNet: An Electronic Lexical Database**. Bradford Books, 1998.

[11] Steven Bird and Edward Loper. NLTK: The natural language toolkit. In **Proceedings of the ACL Interactive Poster and Demonstration Sessions**, pp. 214–217, Barcelona, Spain, July 2004. Association for Computational Linguistics.

## A KL 情報量の近似式

KL 情報量が十分に小さいとき近似的に対称になり $\mathrm{KL}(p(\cdot|w) \parallel p(\cdot)) \simeq \mathrm{KL}(p(\cdot) \parallel p(\cdot|w))$ である．これに $p(w'|w) = \nu p(w') e^{\langle u_w, v_{w'} \rangle}$ を代入すると，

$$\begin{aligned}
&\mathrm{KL}(p(\cdot) \parallel p(\cdot|w)) \\
&= \sum_{w' \in V} p(w') \log \frac{p(w')}{p(w'|w)} \\
&= \sum_{w' \in V} p(w') \Big( -\log \nu - \langle u_w, v_{w'} \rangle \Big) \\
&= -\log \nu - \langle u_w, \bar{v} \rangle
\end{aligned}$$

ところで，$\sum_{w' \in V} p(w'|w) = 1$ であることから，

$$\begin{aligned}
1 &= \sum_{w' \in V} p(w'|w) \\
&= \sum_{w' \in V} \nu p(w') e^{\langle u_w, v_{w'} - \bar{v} \rangle} e^{\langle u_w, \bar{v} \rangle} \\
&= \nu e^{\langle u_w, \bar{v} \rangle} \sum_{w' \in V} p(w') \Big( 1 + \langle u_w, v_{w'} - \bar{v} \rangle + \\
&\qquad \tfrac{1}{2} \langle u_w, v_{w'} - \bar{v} \rangle^2 + \cdots \Big) \\
&\simeq \nu e^{\langle u_w, \bar{v} \rangle} \Big( 1 + \langle u_w, \bar{v} - \bar{v} \rangle + \tfrac{1}{2} u_w^\top \mathrm{Cov}(v) u_w \Big) \\
&\simeq \nu e^{\langle u_w, \bar{v} \rangle} \exp\Big( \tfrac{1}{2} u_w^\top \mathrm{Cov}(v) u_w \Big)
\end{aligned}$$

ただしテイラー展開の高次項は無視している．この式を KL 情報量の式に代入すると

$$\mathrm{KL}(w) \simeq \tfrac{1}{2} u_w^\top \mathrm{Cov}(v) u_w$$

## B 実験設定

**■共起行列の計算** コーパス長 $N = 170\mathrm{M}$ の text8 コーパス[2]を用いる．単語の共起行列 $(n_{w,w'})_{w,w' \in V}$ の計算では $h = 10$ とし，コーパスの端点を除いて前後 $\pm h$ の単語を共起した単語として数える．

**■単語集合 1** 単語の KL 情報量，頻度，単語ベクトルの長さの関係を調べる実験（図 1, 図 2, 図 4）では，コーパスの最初の 200 単語と，出現頻度順にソートして 50 単語おきに選んだ 1000 単語の合計 1200 単語を用いる．

**■単語集合 2** 単語の意味の強さを確認する実験では，単語の品詞として固有名詞 (proper nouns)，機能語 (function words)，動詞 (verbs) を用いる．固有名詞は English Proper nouns データベース[3]に掲載されている 61711 単語の中から text8 コーパスに登場する 10561 単語を使用する．機能語と動詞は NLTK[11] の POS tagging によって取得する．機能語は POS tagging によって{IN, PRP, PRP$, WP, WP$, DT, PDT, WDT, CC, MD, RP}でタグ付けされた単語の中から text8 コーパスに登場する 123 単語を使用する．動詞は POS tagging によって{VB, VBD, VBG, VBN, VBP, VBZ}でタグ付けされた単語の中から text8 コーパスに登場する 4771 単語を使用する．

2） http://mattmahoney.net/dc/textdata.html
3） https://github.com/jxlwqq/english-proper-nouns/