日本語Wikipediaのword2vec表現と語彙特性との関係

○近藤公久[1]・浅川伸一[2] （[1]工学院大学・[2]東京女子大学）

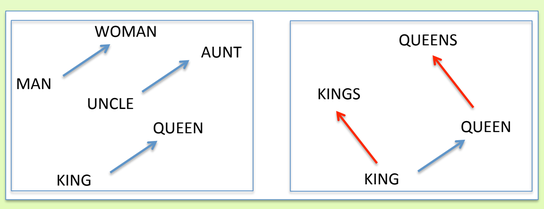
■はじめに

近年，単語埋め込みモデル (word embedding) または ベクトル空間モデル (vector space) と呼ばれる一連の機械学習手法(Mikolov et al., 2013)が意味を表現する可能性があることが示されて注目されている。

単語埋め込みモデルとは，自然言語処理において各々の単語は総語彙数長を次元とする疎表現であるワンホットベクトル(onehot vector)を密表現ベクトル (dense) に変換するために考案された中間表現に由来する(Elman, 1990)。

単語埋め込みモデルで表現された空間上のベクトル間では加算，減算などの演算が形式的には定義できる ことから意味の表現の候補とみなしうる。

例）
vec(王)-vec(男)+vec(女)=vec(女王) (下図参照)は人間の類推機能と比肩しうる。



■目的

NTT データベースシリーズに収録されている単語親密度および単語頻度(天野と近藤, 1999, 2000)，単語心像性(佐久間ら,2008)などの語彙特性は，日本語の様々な特性を反映した特性値である。単語親密度は35名の被験者による主観的評定値であり，人の各単語に対する「なじみ」の程度を表している。単語頻度は新聞12年分をコーパスとして単語の出現回数を数えることによって作成されたものである。また単語心像性は，単語親密度と同様に主観的評定値であり，各単語に対する意味のイメージしやすさを表している。一方，word2vec（e.g., Mikolov, et al., 2013)を用いて構築されるモデルは，膨大な量のテキストを学習することで，単語の共起や構文、そして，意味的空間を表現可能であると考えられている。本研究は，語彙特性データベースの値とword2vecによって学習された単語のベクトル空間表現との関係を解析することを目的とした。すなわち，word2vecのベクトル空間は何を学習し，何を表現しているのかを明らかにすることを意図し，心理実験から得られたNTTデータベースの情報と日本語ウィキペディアの単語埋め込みモデルとを比較することで，両データで表現された情報の関連の検討を試みた。

■方法

コーパス： 日本語Wiki
(http://dumps.wikimedia.org/jawiki/latest)
＊ 12/sep/2016 にダウンロード
＊ MeCabによって形態素（単語）に分割

コーパスサイズ：
MeCabの出力1,265,697,011形態素(分割素片)。出現回数が5回未満の語を除いた異なり数で1,140,357 (vocabularies), トータルで5,774,139トークン。

＊Word2vecにより学習
Word2vecの次元は128。ウィンドウ幅 10, 負事例サンプリング 10, skipgram を用いた。

解析対象：
wikipediaに出現する単語とNTTデータベース(以降NTT-DBと呼ぶ)に存在する単語とのマッチングがとれた17,842語。

■結果（重回帰分析）

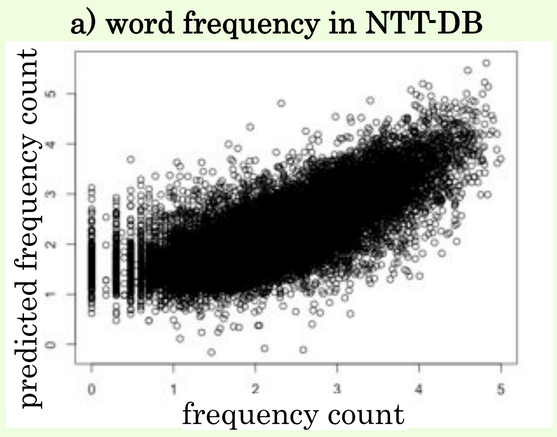
17,842語のword2vecパラメータ(128個）による各語彙特性値の回帰モデル

	adjusted-R^2
NTT-DB単語頻度（対数）	0.5495
NTT-DB単語心像性（文字）	0.4232
NTT-DB単語親密度（文字）	0.3741
Wikipedia内の単語出現頻度（対数）	0.6539
NTT-DB単語心像性（音声）	0.4242
NTT-DB単語親密度（音声）	0.3189

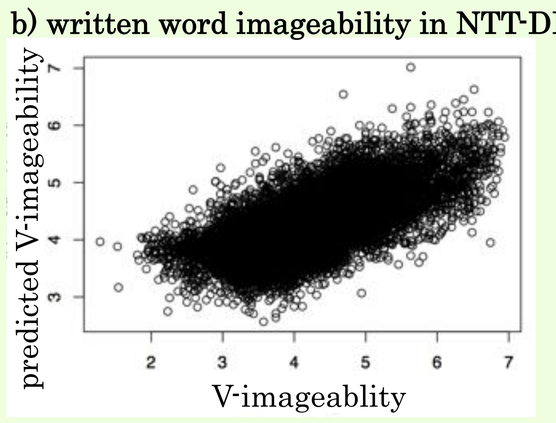
■考察

word2vecにより学習した各単語のベクトル値による語特性の推定は，頻度に対して高成績である。一方，主観的な評定値である親密度及び心像性に対しては頻度の推定より劣る。両者を比較すると、若干ながら心像性の方が精度が高い。心像性が意味のイメージしやすさを示す特性であることから，word2vecのパラメータ空間が意味と関連することと関係していると考えられる。
また，頻度の推定おいては，学習に用いたWikiにおける頻度の方が高成績であり，NTT−DBが新聞に出現する頻度であるという特性の違いが現れている。また，文字単語親密度に比べて音声単語親密度の推定成績が悪い。これは文字（漢字）そのもの難しさなどの特性が関与しているためと考えられる。一方，心像性は入力メディアによる違いは小さく，両特性の文字単語と音声単語による意味のイメージしやすさに大きな違いがないためと考えられる。

a) word frequency in NTT-DB



b) written word imageability in NTT-DB



c) written word familiarity in NTT-DB

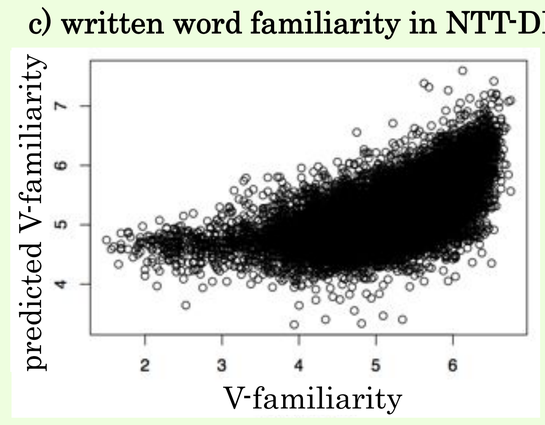
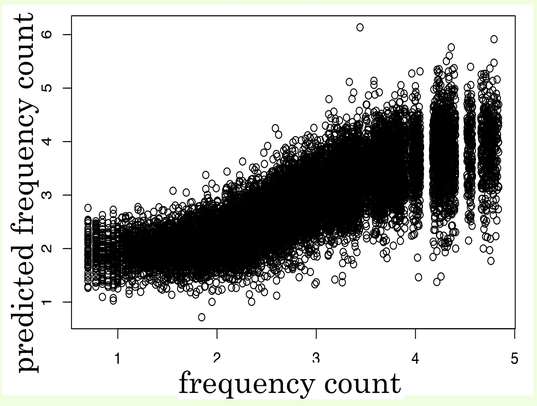
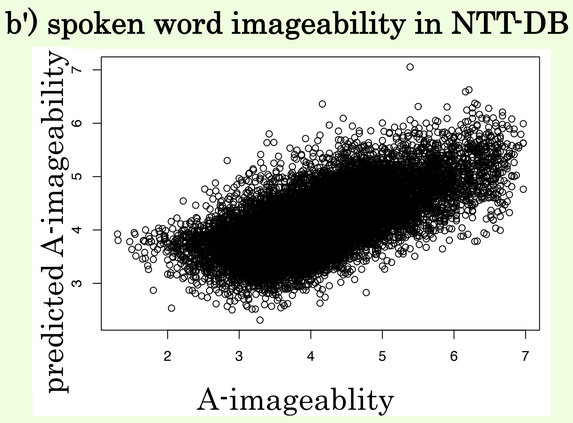


Fig. 1 各特性値と推定値の相関図

a') word frequency in Wiki



b') spoken word imageability in NTT-DB



c') spoken word familiarity in NTT-DB

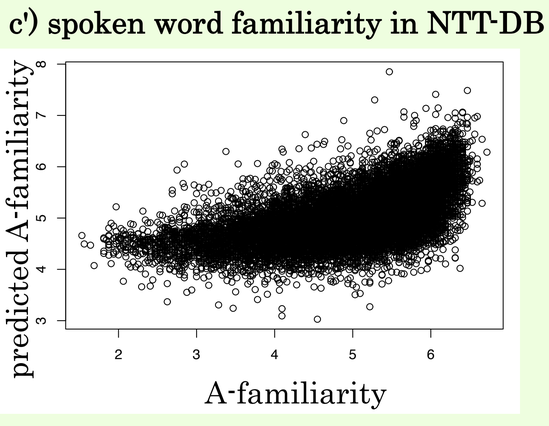


Fig. 1' 各特性値と推定値の相関図

■demo サイト

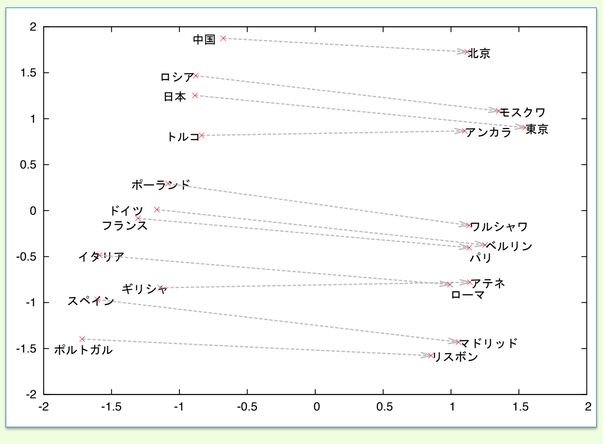
本発表で用いたword2vec学習済みデータ（vector data），およびpython (jupyter notebook)によるword2vecのサンプルコード:
https://github.com/ShinAsakawa/2017jpa/
ただしデータは 2017年7月の日本語ウィキペディア。次元数 200，300，の二種。アルゴリズムは skipgram, および CBOW。

■今後の予定

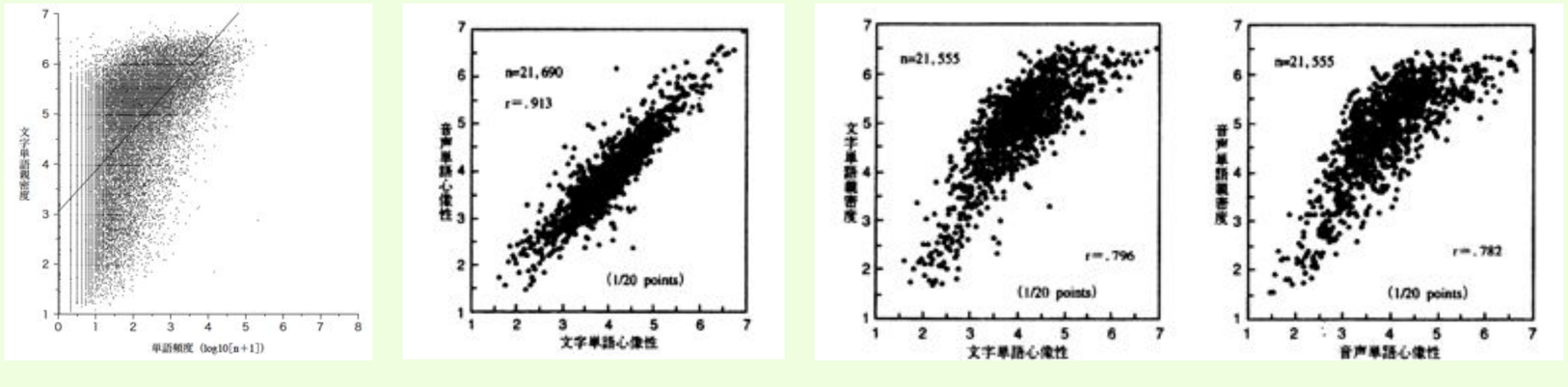
今後さらにモデルの解析するを進め，学習されたベクトルがどのような役割を果たしているのかを探求する予定である。 推定精度を挙げることで，新語が出現したり時代変化する主観評定による指標を自動生成することを試みる。

■参考文献

天野成昭・近藤公久 (1999, 2000). 日本語の語彙特性第1巻単語親密度，第7巻単語頻度，三省堂.
佐久間尚子・伊集院睦雄・伏見貴夫・辰巳格・田中正之・天野成昭・近藤公久 (2008). 日本語の語彙特性第8巻単語心像性，三省堂.
Mikolov, T., Yih, W. tau, & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), Atlanta, WA, USA.
他



APPENDIX A
word2vecで学習したvector値の2次元による国と首都の関係（Mikolov et al.,2013 による）



APPENDIX B NTT-DB収録の頻度，親密度，心像性間の相関
(天野，近藤，1999; 2000；佐久間ら, 2008より)