LIWC2001手作業翻訳の方針と半自動翻訳手法の提案

山本眞大[1] 那須川哲哉[2] 上條浩一[2] 北村英哉[3]

[1]慶應義塾大学大学院 理工学研究科
[2]日本アイ・ビー・エム株式会社 東京基礎研究所
[3]関西大学 社会学部

1 はじめに

LIWC (Linguistic Inquiry and Word Count) [1]とは、語彙を抽象化してカテゴリ化するためのツールであり、テキスト分析の際に使用される。例えば、LIWC2001には“Seeing”や“Cognition”などの計74個のカテゴリがあり、各カテゴリに属する単語が定義されている。LIWCのカテゴリは選挙運動の分析や著者の年齢推定、性格推定などの様々なタスクの素性として使用されている[2–4]。しかしながら、LIWCの日本語版は存在しないため、日本語で書かれた文章に対してLIWCを適用することは困難である。

そこで我々は、心理学的知見と整合性を取りつつ、LIWC2001を手作業で日本語に翻訳した。本論文では、まず手作業翻訳の概要について述べる。次に、手作業翻訳により得られた知見を基にした、LIWC2001の半自動翻訳手法を提案する。

評価実験では、半自動で翻訳されたLIWC2001を用いた性格推定効果の測定が行われた。その結果、提案手法の有効性が示唆された。

2 LIWC2001

LIWC2001とは、テキスト分析の際に用いられるツールである。LIWC2001には、“Word Count”や“Seeing”、“Cognition”等、計74個のカテゴリがある。それらのうち6カテゴリは、“Word Count”や“Words per sentence”等の辞書による定義が必要のないカテゴリである。残りの68個のカテゴリには、そのカテゴリに属する単語が定義されている。例えば、“Seeing”カテゴリであれば、*“view”*や*“see”*、*“look”*などの単語が定義されている。本論文では、その68個のカテゴリに属する単語の翻訳について述べる。

3 LIWC2001の手作業翻訳

本章では、LIWC2001の手作業翻訳の方針について述べる。基本的には、LIWC2001で辞書化されている68カテゴリのうち、日本語には存在しないカテゴリ等を除く66のカテゴリに対応する日本語表現を定義し、IBM Watson Explorer Advanced Edition Analytical Components V11.0 （以下WEXと略記）[5]の言語処理機能を利用して、テキスト中の各表現の出現を特定する仕組みを実装する。実装においては基本的に、自立語の単語表現は辞書登録を行い、付属語を含む複合表現や否定形の判断が必要となる表現などは形態素解析結果や構文解析結果を利用するWEXのパタン抽出機能を用いる。

具体的には、まず、以下の通りにLIWC2001の各カテゴリについて翻訳の方針をたてた。

- 削除カテゴリ群: 日本語には存在しないか、翻訳すると内容が重複するため削除するカテゴリ。
- Wikipedia利用カテゴリ群: Wikipediaを利用して単語を追加するカテゴリ。
- 構文パタン作成カテゴリ群: 構文パタンを作成するカテゴリ。
- 翻訳辞書使用カテゴリ群: 翻訳辞書による翻訳結果からノイズとなり得る単語を削除するカテゴリ。
- まとめカテゴリ群: 各カテゴリの上位カテゴリ。

表1に各カテゴリ群に対応するカテゴリを示す。

3.1 削除カテゴリ群

日本語とは対応が取れない“Article (冠詞カテゴリ)”を削除した。また、“Fillers (フィラーカテゴリ)”は“Nonfl (会話に影響を与えない単語を集めたカテゴリ)”と内容が重複するため削除した。

表 1: 各カテゴリ群に対応するカテゴリ

カテゴリ群	LIWC2001 カテゴリ
削除カテゴリ群	Article, Fillers
Wikipedia 利用カテゴリ群	I, We, You, Other,Othref
構文パタン作成カテゴリ群	Negate, Preps, Number, Past, Present, Future, Swear, Nonfl
翻訳辞書使用カテゴリ群	Assent, Posfeel, Optim, Anx, Anger, Sad, Cause, Insight, Discrep, Inhib, Tentat, Certain, See, Hear, Feal, Comm, Friends, Family, Humans, Time, Up, Down, Incl, Excl, Motion, School, Job, Achieve, Home, Sports, TV, Music, Money, Relig, Death, Body, Sexual, Eating, Sleep, Groom
まとめカテゴリ群	Pronoun, Self, Affect, Posemo, Negemo, Cogmech, Senses, Social, Space, Occup, Leisure, Metaph, Physcal

3.2 Wikipedia 利用カテゴリ群

対応する Wikipedia のページから単語が追加される。例えば、“1st person singular (一人称代名詞カテゴリ)” に属する単語は、Wikipedia の「日本語の一人称代名詞」のページ [1] に記載されている単語が追加される。しかし、上記には、「朕」や「麻呂」などの使用頻度が非常に少ない単語が存在する。そういった単語を追加するとノイズになる可能性がある。そこで、常識的に使用頻度が少ないと考えられる単語は削除した。

3.3 構文パタン作成カテゴリ群

LIWC2001 に収録されている単語を翻訳する必要性がないため、構文パタンを作成するカテゴリである。例えば “Present (現在形の単語を集めたカテゴリ)” では、「〜する。」等の現在形を特定する構文パタンを作成し、WEX のパタン抽出機能を用いて出現を特定する。但し、“Swear” カテゴリの出現の特定には、文献 [6] の「軽卑表現」に相当するパタンを用い、“Preps (前置詞カテゴリ)” には助詞を対応させた。

3.4 翻訳辞書使用カテゴリ群

各英語表現に対して複数の日本語候補を収録した英日機械翻訳用の対訳辞書を用いて対応するカテゴリに属する単語を翻訳した。しかし、この段階ではノイズとなり得る単語が多数存在する上に、翻訳辞書中に存在しない単語は追加されない。そこで、単語の削除および単語の追加を行った。

単語の削除

以下の基準に合致する単語を削除した。

- **使用頻度が低いと考えられる単語**
- **カテゴリの意味に明らかにそぐわない単語**
  常識的に考えてカテゴリの意味に明らかにそぐわない単語を削除した。例えば、“Seeing” カテゴリには *“saw”* という単語が存在するので、「のこぎり」という単語が翻訳結果として出力される。そのような明らかにカテゴリの意味と異なる単語を削除した。
- **カテゴリの意味で使われることが少ない単語**
  実際の使われ方がカテゴリの意味と異なる単語を削除した。具体的には、twitter をランダムサンプリングして収集したコーパス (以下 twitter コーパスと略記) を用いて、各単語がどのような意味で使われているかを調査し、カテゴリの意味で使われていない単語は削除した。例えば、“Seeing” カテゴリには *“eye”* という単語が存在するので、「目」という単語が翻訳結果として出力される。しかしながら、「目」という単語は「5 年目」や「目が痛い」等の「見る」という意味で使われない事が多い。よって、そのような単語を削除した。

単語の追加

翻訳辞書に存在しない単語の追加を行った。例えば、“Seeing” カテゴリには「見える」や「観る」等の単語の追加を行った。

3.5 まとめカテゴリ群

各カテゴリの上位カテゴリとみなすことが出来るカテゴリ群である。**3.1**〜**3.4** にて各カテゴリに属する単語群が決定した後、その和集合が対応するカテゴリに属する単語となる。

4 LIWC2001 の半自動翻訳

本章では LIWC2001 の半自動翻訳手法について述べる。手作業翻訳と同様に、5 つの方針で半自動翻訳を行う。半自動翻訳の際には、**3.1**〜**3.3**、**3.5** で述べた方法はそのままに、**3.4** で説明した翻訳辞書を使用するカテゴリに属する単語の削除を自動化する。

まず、英日対訳辞書とのマッチングにより、**3.4** で説明したカテゴリに属する単語が自動で翻訳される。次に以下の要領でノイズとなり得る単語が自動で削除される。

[1] https://ja.wikipedia.org/wiki/日本語の一人称代名詞

4.1 カテゴリ代表語の決定

まず、各カテゴリの代表語を決定する。各カテゴリの代表語の決定は以下の手順により行なわれる。

1. カテゴリ cat_i のカテゴリ平均ベクトル $\mathbf{v}_{cat_i}$ を以下の式により求める。

$$\mathbf{v}_{cat_i} = \frac{\sum_{j=1}^{n_{cat_i}} \mathbf{v}_{w_j}}{n_{cat_i}} \tag{1}$$

ここで n_{cat_i} は、カテゴリ cat_i の翻訳後の単語数であり、$\mathbf{v}_{w_j}$ は翻訳後の単語 w_j の分散表現である。各単語の分散表現は、twitter コーパスに対して文献 [7] の手法を適用して取得される。なお、1 ツイートを 1 文と見なし、分かち書きの際には WEX の形態素解析器を使用した。

2. カテゴリ平均ベクトル $\mathbf{v}_{cat_i}$ と単語ベクトル $\mathbf{v}_{w_j}$ のコサイン類似度 $cos(\mathbf{v}_{cat_i}, \mathbf{v}_{w_j})$ が以下の式により求められる。

$$cos(\mathbf{v}_{cat_i}, \mathbf{v}_{w_j}) = \frac{\mathbf{v}_{cat_i} \cdot \mathbf{v}_{w_j}}{||\mathbf{v}_{cat_i}|| \cdot ||\mathbf{v}_{w_j}||} \tag{2}$$

3. 単語 w_j が翻訳結果として現れた回数を n_{w_j} とする。
4. 単語 w_j のスコア $Score_{w_j}$ が以下の式により求められる。

$$Score_{w_j} = cos(\mathbf{v}_{cat_i}, \mathbf{v}_{w_j}) \cdot n_{w_j} \tag{3}$$

5. $Score_{w_j}$ の一番高い単語がカテゴリ代表語となる。
6. 上記 1〜5 が対応するカテゴリ全てに対して行なわれる。

4.2 カテゴリの意味に明らかにそぐわない単語の削除

カテゴリ代表語が決定した後、カテゴリの意味にそぐわない単語が削除される。以下の手順により、単語の削除が行なわれる。

1. カテゴリ代表語の分散表現と翻訳後の各単語の分散表現とのコサイン類似度が計算される。
2. 類似度が δ 未満の単語が削除される。
3. 上記 1〜2 が対応する全てのカテゴリに対して行なわれる。

4.3 カテゴリの意味で使われることが少ない単語の削除

次に、カテゴリの意味で使われることが少ない単語が削除される。以下の手順により行なわれる。

1. twitter コーパスから、カテゴリ代表語を含む文 s_{rep} が N_{rep} 件ランダムに抽出される。
2. カテゴリ代表語の文脈ベクトル $\mathbf{v}_{s_{rep}}$ が以下の式により求められる。

$$\mathbf{v}_{s_{rep}} = \frac{\sum_{i=1}^{N_{rep}} \mathbf{v}_{s_{rep_i}}}{N_{rep}} \tag{4}$$

ここで、$\mathbf{v}_{s_{rep_i}}$ は、s_{rep} 中の文 s_{rep_i} のカテゴリ代表語前後 5 単語の分散表現である。

3. 単語 w_j を含む文 s_{w_j} が N_{w_j} 件ランダムに抽出される。
4. 単語 w_j の文脈ベクトル $\mathbf{v}_{w_j}$ が以下の式により求められる。

$$\mathbf{v}_{s_{w_j}} = \frac{\sum_{i=1}^{N_{w_j}} \mathbf{v}_{s_{w_{j_i}}}}{N_{w_j}} \tag{5}$$

ここで、$\mathbf{v}_{s_{w_{j_i}}}$ は、s_{w_j} 中の文 $s_{w_{j_i}}$ のカテゴリ代表語前後 5 単語の分散表現である。

5. $\mathbf{v}_{s_{rep}}$ と $\mathbf{v}_{s_{w_j}}$ のコサイン類似度 $cos(\mathbf{v}_{s_{rep}}, \mathbf{v}_{s_{w_j}})$ が求められる。
6. $cos(\mathbf{v}_{s_{rep}}, \mathbf{v}_{s_{w_j}})$ が大きい順に γ_{cat_i} 個の単語が抽出される。
7. 上記 1〜6 が対応するカテゴリ全てに対して行なわれる。

5 評価実験

提案手法の有効性を示すために評価実験を行った。実験では、半自動で構築された LIWC2001 を用いて性格推定効果の測定を行った。ここで、性格推定効果の測定の際には文献 [8] の手法を用いた。また、比較対象として以下を用いた。

- ベースライン: 学習データ中の平均値を推定値とする手法。
- 人手: 人手により構築された LIWC2001 を用いる手法。

5.1 実験条件

twitter コーパスとして、2012 年 7 月 1 日〜2012 年 9 月 31 日の期間にランダムサンプリングして取得した計 4,736,397 ツイートを用いた。文献 [7] の手法により単語の分散表現を取得する際の各種パラメータは次元数 200、CBOW、窓幅 5、最低頻度を 5 とした。**4.2** で用いられる類似度閾値 δ は 0.05 に設定した。また、**4.3** で抽出される単語の個数 γ_{cat_i} は、カテゴリ

表 2: 性格の推定値と実値の MAE

Profile		提案	人手	ベースライン
Big5	Extraversion	0.121	0.117	0.129
	Agreeableness	0.119	0.114	0.127
	Conscientiousness	0.111	0.109	0.114
	Neuroticism	0.135	0.135	0.146
	Openness	0.124	0.123	0.132
Needs	Structure	0.091	0.091	0.098
	Practical	0.105	0.105	0.109
	Challenge	0.146	0.139	0.164
	Self	0.126	0.123	0.136
	Excitement	0.146	0.143	0.158
	Curiosity	0.120	0.112	0.123
	Liberty	0.111	0.107	0.116
	Ideal	0.132	0.126	0.137
	Harmony	0.084	0.085	0.088
	Love	0.152	0.148	0.166
	Close	0.122	0.117	0.128
	Stable	0.104	0.104	0.110
Values	Self transcendence	0.091	0.090	0.095
	Openness to change	0.118	0.111	0.124
	Hedonism	0.137	0.135	0.146
	Self enhancement	0.136	0.136	0.147
	Conversion	0.094	0.091	0.097
Average		0.119	0.116	0.127

cat_i に属する翻訳前の単語の個数とした。実験の際には、文献 [8] の方法で収集された 249 名の twitter データを用いた。ツイート数の合計は 213,068 ツイートである。

5.2 実験結果・考察

表 2 に各手法による性格の推定値と実値の Mean Absolute Error (MAE) を示す。ここで、MAE の計算においては、0 から 1 の値を取る性格スコアに対して重回帰分析を行い、10-分割交差検証により性格の推定値と実値の MAE を測定した。表 2 より、提案手法により構築された LIWC2001 を用いた性格の推定値と実値の MAE の平均は、ベースラインの MAE の平均よりも小さいことが分かる。このことから、提案手法により半自動で構築された LIWC2001 の有効性が示唆される。

一方、提案手法により構築された LIWC2001 を用いた性格の推定値と実値の MAE の平均は、人手により構築された LIWC2001 を用いた性格の推定値と実値の MAE の平均よりも大きいことが分かる。この原因としては大きく分けて以下の 2 点が考えられる。

まず第一に、単語の削除に失敗し、ノイズとなり得る単語が残ってしまったことが挙げられる。特にカテゴリ代表語の決定に失敗するとノイズとなり得る単語が残り易い傾向があった。第二に、半自動で翻訳する際には、単語の追加が行われなかったことが挙げられる。上記の問題に対しては、カテゴリ代表語の決定方法の改善、単語の追加方法の検討を行なうことで対処することが考えられる。

6 おわりに

本論文では、LIWC2001 の手作業翻訳の方針と半自動翻訳手法について述べた。まず、大きく分けて 5 つの方向性で LIWC2001 を手作業で翻訳した。そして、翻訳辞書を用いるカテゴリについては、ノイズとなり得る単語の削除を自動化する手法を提案した。評価実験では、半自動で翻訳された LIWC2001 を用いた性格推定効果の測定が行われた。その結果、提案手法の有効性が示唆された。

今後は、翻訳辞書に存在しない単語を自動で追加する方法等についても検討していく予定である。

参考文献

[1] James W Pennebaker et al. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, Vol. 71, pp. 1–21, 2001.

[2] Andranik Tumasjan et al. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, Vol. 10, pp. 178–185, 2010.

[3] Dong Nguyen et al. Author age prediction from text using linear refresion. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pp. 176–178, 2011.

[4] Hernan Badenes et al. System u: Automatically deriving personality traits from social media for people recommendation. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pp. 373–374, New York, NY, USA, 2014. ACM.

[5] Wei-Dong Zhu et al. Ibm watson content analytics: Discovering actionable insight from your content. *An IBM Redbooks publication. ISBN-10:0738439428*, 2014.

[6] 荻野紫穂, 那須川哲哉, 金山博, 榎美紀. 軽卑表現の情報を活用した知識発見. 言語処理学会第 18 回年次大会予稿集, pp. 58–61, 2012.

[7] Tomas Mikolov et al. Distributed representations of words and phrases and their compositionality. *NIPS*, Vol. 10, pp. 178–185, 2013.

[8] 那須川哲哉, 上條浩一, 山本眞大, 北村英哉. 日本語における筆者の性格推定のための言語的特徴の調査. 言語処理学会第 22 回年次大会予稿集, p. To appear, 2016.