多言語極性辞書の構築とその包括的評価

岩本 蘭
慶應義塾大学　理工学研究科
r.iwamoto@keio.jp

金山 博
日本アイ・ビー・エム株式会社 東京基礎研究所
hkana@jp.ibm.com

1　はじめに

本論文では言及単位の評価表現抽出のための多言語極性辞書の作成および評価方法について述べる．始めに活用形の違いを吸収できる lemma ベースの英語辞書を機械翻訳と分散表現を用いて多言語に拡張し，次に辞書作成の際に言語ごとに加える処理を議論する．最後に多言語辞書の性能を包括的に測るための評価手法を述べる．

英語の極性辞書に関する研究は synset ごとの極性スコアを利用した SentiWordNet を用いた手法 [1, 2] をはじめとして数多く存在する．また，分散表現を利用した対訳辞書構築 (Bilingual Lexicon Induction: BLI) [3] は極性辞書の多言語拡張手法としても用いることができる．Huang の研究 [4] では分散表現で同じ語が何回も類似語として選択される現象 (Hubness) を軽減する手法を考案し，元の極性表現からより様々な語を抽出できるようにした．このように極性辞書の 1-2 言語への拡張を行った研究はいくつか存在するが，多言語の極性辞書を統一基準で作成した研究は少ない．

分散表現を用いない辞書拡張の手法として 2 言語分の辞書をもとにして他言語の辞書を作成する研究 [5] や，パラレルコーパスや多言語のコンセプトグラフを活用する研究 [6, 7] がある．極性辞書に関する研究の多くは辞書の活用先として言及単位の評判分析を想定しておらず，語彙項目には単語の表層形や極性スコアの情報しか付与されていない．そこで我々は構文構造を活用した言及単位の評判分析を実現すべく，活用形の違いを吸収できる lemma ベースの辞書を作成した．

2　多言語極性辞書作成

2.1　極性辞書

言及単位の評判分析では与えられた文に対して構文解析を行い，評価表現を検出したい節中の lemma が極性辞書と符合したときに，極性と述語，その対象を同定する．極性辞書は表 1 のような形式になっている．辞書には形容詞，副詞，動詞と名詞が語彙として含まれる．本研究では格フレームの情報を付与した英語の極性辞書 [8] をベースに拡張したものを用いる．

表 1: 極性辞書の例．lemma，品詞，格フレームの組に対し好評 (+) または不評 (−) の極性を割り当てる．格フレーム中の下線は対象となる格を示す．

lemma	品詞	極性	格フレーム
friendly	ADJ	+	obj
love	VERB	+	nsubj, obj
unfortunately	ADV	−	
noise	NOUN	−	

辞書の語彙項目は表層形ではなく lemma ベースで記述した．極性表現が様々な活用形を持つときにも対応でき，辞書の記述を簡潔にするためである．

2.2　極性辞書の多言語拡張

本論文では英語の極性辞書を 16 言語に拡張した．翻訳と分散表現をそれぞれ用いて辞書を作成しそれらを統合することによりカバレッジが向上した．辞書の作成方法について以下に述べる．

翻訳を用いた辞書作成　英語の語彙項目のうち形容詞と動詞を他言語に翻訳するために Watson Language Translator[1]を用いた．準備段階として英語の極性語を含むテンプレート文を作成する．例えば “He is [ADJ].”,“We can [VERB].”, “We can [VERB] XYZ.” という文を作成し，[ADJ], [VERB] の部分を語彙項目の lemma に置き換える．表 1 の “love” のように，格フレームに obj を持つ動詞には架空の目的語 ‘XYZ’ を付加する．翻訳後の文を Universal Dependencies (UD) [9] の構文解析器を用いて構文解析する．他の語彙項目を代入した文との共通部分のうち，数が多い 1-2 例を “He is”, “We can” の翻訳結果だと仮定し，正しく翻訳されているとみられる文のみから語彙項目を抽出する．その後 [ADJ], [VERB] の部分の翻訳結果の lemma を辞書に追加する．ただし言語ごとに構文解析器の性能に偏りがあり，正しい lemma が出力されないことがあるため，なるべく翻訳後の表層形と lemma が等しくなるような英文を用いた．

[1] https://lanugage-translator-demo.ng.bluemix.net/

表 2: 極性辞書の語彙数．翻訳と分散表現を用いて拡張した語彙数とそれらを統合した辞書の語彙数を翻訳, 分散表現, 合計に示す．‘-’ は語彙資源が存在しないことを表す．英語は人手で作成した辞書を用い，日本語は翻訳を用いて作成した辞書に人手で作成した辞書を加えた．

言語	提案法 翻訳	分散表現	合計	MSL
英語 (en)			3385	1727
アラビア語 (ar)	499	409	820	1173
チェコ語 (cs)	1491	1051	2052	2138
ドイツ語 (de)	1906	879	2399	1913
スペイン語 (es)	1665	992	2001	2332
フィンランド語 (fi)	1101	593	1393	1390
フランス語 (fr)	1375	961	1926	2652
ヘブライ語 (he)	490	549	880	1169
インドネシア語 (id)	641	520	921	1121
イタリア語 (it)	1512	942	1902	2284
日本語 (ja)	385	-	2080	225
韓国語 (ko)	584	-	584	621
オランダ語 (nl)	1030	842	1521	1887
ポルトガル語 (pt)	1787	849	2082	1791
ロシア語 (ru)	1374	1003	1947	2340
トルコ語 (tr)	286	436	664	653
中国語 (zh)	737	-	737	96

形容詞や動詞の表層形と lemma が同形となる例として，多くのヨーロッパ言語では男性単数形の主語と be 動詞と共に形容詞を用いたときや複数形の主語と助動詞と共に動詞を用いたときがある．また，アラビア語やヘブライ語の動詞は男性単数の過去形 (完了形) の時に表層形が lemma と同形になる．翻訳前の英語の格フレームをそのまま翻訳先の言語の辞書でも用いるため，翻訳後の文を構文解析し，極性語の品詞が翻訳前の品詞と一致した時のみ辞書に追加した．

分散表現を用いた辞書作成　次に Multilingual Word Embeddings(MUSE) [10] の訓練済みの多言語分散表現を用いて極性辞書を 1 言語ずつ作成した．まずは対象言語の分散表現の中から，英語の極性辞書にある lemma との cos 類似度が高い順に単語ベクトルを 5 つ取得する．選んだ単語の品詞を UD コーパスを用いて同定する．MUSE では単語は表層形のまま記述され品詞の区別はしていない．そこで UD コーパスからその単語の表層形を探し，品詞と lemma を推定した．選んだ単語の中で品詞が元の英単語と一致していて最も類似度が高い単語を 1 つだけ辞書に追加した．

分散表現では反対の極性を持つ単語の類似度が高くなる問題が指摘されている [11, 12]．しかし分散表現を用いて英語の極性語を他言語に拡張するときには，英語の極性語と同じ品詞を持つ中で類似度が最も高い他言語の単語の多くは，英語の極性語と同じ極性を持つと経験的に分かった．また，条件を満たし類似度の高い単語を複数個辞書に追加すると反対の極性の単語も誤って辞書に追加される可能性が高くなるため，1 単語のみを辞書に追加した．

また大規模コーパスを用いて作成された分散表現には頻度が比較的低い単語も含まれている．そこで UD コーパスを用いて表層形が一致する単語の品詞や lemma を割り当てることで頻度のフィルタリングを暗に行なっている．しかし言語によって UD コーパスのサイズが大きく異なる．例えばインドネシア語は他の言語と比べてコーパスのサイズが小さいため，分散表現で獲得できる語彙数は多くなかった，

語彙数比較　翻訳と分散表現からそれぞれ辞書を作成しそれらを統合した．語彙項目が翻訳由来と分散表現由来で矛盾した極性を持つ場合は両方とも辞書から削除した．否定の副詞や，“think” など評価表現抽出の際にその子ノードを探索するために用いる単語を辞書から削除した．作成した多言語辞書の語彙数を表 2 に示す．比較手法として機械翻訳から知識グラフを作成し極性表現を集めた Multilingual Sentiment Lexicon [13] (以降 MSL と記す) を用いた．MSL では極性語の表層形が提供されており，我々の辞書と形式を等しくするために UD コーパスを用いて MSL の極性語を lemma に変換し格フレームを付与した．

翻訳と分散表現を用いた方法の 2 つを組み合わせることでカバレッジが上がっていることがわかる．分散表現では英語の単語は同じ意味を持つ頻度の高い単語に変換され，翻訳では英語に似た表現に変換される傾向がある．ドイツ語を例にとると，分散表現から作成した辞書には “schlimm”(悪い) や “böse”(怒っている) などドイツ語でよく使われるが英語の単語と綴りが似ていない語が多く含まれるのに対し，翻訳から作成した辞書には英語と使い方が似ている “über-” (“over-”) や “-voll” (“-ful”) などの接頭辞や接尾辞を持つ単語や英語の綴りと似ている語が多く含まれている (“respektvoll”, “wundervoll” など)．

2.3　言語ごとの追加処理

UD では lemma についての明確な定義を設けておらず，各言語に委ねられている．ここでは UD での多言語の lemma の定義の違いに触れ，極性辞書を作成する際の注意点について述べる．

表記ゆれへの対応　アラビア語ではコーランや子供向けの書籍には母音の読みを示した Tashkil という発音記号が振られている．UD のアラビア語コーパスでの

表 3: 整形後のデータセット．+,– は好評/不評表現を含む文の数，文長は一文中の語数の平均値である．

言語	分野	+	−	文長
ar [14]	ホテル	250	250	29.6
cs [15]	レストラン	250	250	16.5
de [16]	カトラリー	297	62	13.7
en [14]	レストラン	250	250	14.7
es [14]	レストラン	250	250	15.4
fr [14]	レストラン	250	250	16.0
he [17]	ニュース	250	250	14.2
id [18]	レストラン	250	250	10.7
it [19]	ホテル	250	250	15.1
ja [20]	携帯電話	238	295	21.5
ko [21]	映画	250	247	9.4
nl [14]	レストラン	250	250	14.9
pt [22]	本	250	250	22.6
ru [14]	レストラン	250	250	17.3
tr [14]	レストラン	250	250	10.3
zh [14]	携帯電話	253	247	35.1

表 4: 多言語極性辞書での評価表現抽出の性能比較．

	提案法		MSL	
言語	適合率	再現率	適合率	再現率
ar	95.5	36.8	83.5	34.0
cs	88.3	36.6	72.5	32.4
de	94.2	46.8	85.6	56.5
en	92.7	46.8	90.8	45.4
es	90.2	36.6	74.0	38.4
fr	90.0	43.6	76.1	53.6
he	82.0	16.4	76.0	28.4
id	92.7	33.8	83.5	33.0
it	88.3	29.8	80.0	42.8
ja	92.2	33.2	66.7	6.2
ko	89.6	10.9	71.4	7.0
nl	90.8	41.6	73.8	50.6
pt	83.2	32.4	78.1	43.8
ru	90.1	30.4	72.5	39.0
tr	91.2	17.0	64.9	13.0
zh	88.2	24.6	75.7	22.0

lemma には発音記号がついているため，辞書作成の際には lemma から発音記号を除去して符号させる必要がある．また，ウムラウトなどの特殊記号がある言語ではキーボード入力の際に代替文字を用いることがある．例えばドイツ語では “ä” を “ae” と表す．表記ゆれに対応することにより応用タスクの性能が向上する場合がある．

接頭辞・接尾辞　接頭辞・接尾辞を lemma に含めるかどうかは言語ごとに異なる．チェコ語では否定の接頭辞 “ne-” を除去した状態を lemma としている．例えば単語 “nemilý”(美しくない) の lemma は “milý”(美しい) となり，‘Polarity = Neg’ の素性がつけられている．よって表層形に “ne-” がつく単語は極性を反転させる処理を追加する必要がある．しかし単語によっては lemma から接頭辞 “ne-” が除去されないまま出力され，極性誤りの大きな原因となる．

インドネシア語では語根に接辞をつけることで時制・品詞の変化や意味の付加を表現している．例えば基語 “beli”(買う) に動詞の接頭辞としてよく使われる “meN-” を付加した “membeli”(-を買う) や，受動態を表す “di-” を付加した “debeli’ という単語の lemma は，UD では接頭辞が付いたまま表示され，素性 ‘Voice’ がつけられている．また，“baik”(良い) という形容詞はそれを語根として “membaik”(良くなる)，“memperbaiki”(-を修理する) と様々な意味を持つので必ずしも語根のみを考慮すればよいわけでもない．インドネシア語の品詞に依らない意味表現と接辞の関係が極性表現と辞書の語彙項目の符号を難しくしており，評判分析タスクにおいて再現率を低める原因となっている．

上の例のほかにも日本語の “人々” やインドネシア語の “orang-orang”(人々) などの重複語の取り扱い，日本語の語の単位 [23] についても多くの議論が残る．またヘブライ語など現状の UD では lemma に対応していない言語もある．辞書作成の際の誤りは本節で述べた言語ごとの違いに起因するものが多い．多言語の辞書を統一基準で作成することで，分析対象の言語について詳細な知識がなくても，他の言語との比較を通じて辞書の誤りやカバレッジの低さの原因を予測し改善することが容易になった．

3　実験

多言語の極性辞書の性能を比較するため，UD を基とする構文解析をした上で辞書や文法規則を適用し，評価表現抽出を行った [24]．言及単位の評判分析では多言語を同一基準で比較可能なデータセットが存在しないことが問題点としてあげられる．SemEval-2016: Aspect Based Semtiment Analysis [14] では 8 言語のデータセットが作成され，その後チェコ語 [15] などでもその形式に則ったデータセットが作成された．しかし我々が今回辞書を作成した 16 言語と英語の大部分をカバーするような統一基準のデータセットは存在しないため，データセットを整形した．表 3 に整形したデータセットの情報を示す．言語の欄に整形前のデータセットを引用している．既存データセットでは基本

的に言及単位または文単位で極性のアノテーションがつけられている．文単位のアノテーションしか存在しない言語も多いため，データセットの中から好評/不評の片方の極性のみを含む文を抜き出して用いた．

文単位での適合率と再現率を評価したものを表 4 に示す．UD2.4 で訓練済の StanfordNLP [25] を用いた．我々の辞書を用いたシステムでは 16 言語中 10 言語で 90 %以上の適合率を達成している．適合率が比較的低い言語のうち，ヘブライ語や韓国語，ポルトガル語はデータセットのドメインが原因であると考えられる．本や映画のドメインでは極性が反転する語やドメイン特有の評価表現が存在する [26] ため，それらの評価表現を辞書に追加することで性能が向上する．その他の検出誤りは皮肉や暗喩，我々の評価表現抽出の方針とアノテーションの方針のずれによって生じる．

MSL では極性を持たない単語が誤って極性辞書に入っていることが多い．例えばスペイン語では “ir”(‘go’) が不評語のリストに含まれていた．多くの既存データセットには極性表現を含まない中性のラベルがついた文が存在しなかったため，今回の実験データは好評/不評どちらかのラベルを持つ．その結果極性を持たない語が誤って好評/不評と認識されたときには適合率は低くなるが，再現率は高くなる．また MSL では特に日本語や韓国語において単語分割の誤りが多く，再現率が低い原因の 1 つとなっている．

4 おわりに

翻訳と分散表現を用いて多言語の lemma ベースの極性辞書を作成し，辞書作成の際に得た言語ごとの lemma に関する知見を紹介した．また多言語評価表現抽出のための評価データを作成し辞書の性能比較を行い，提案法の辞書が高い適合率を達成した．今後は多言語処理の知見をさらに多くの言語や他の応用タスクに適用していきたい．

参考文献

[1] A. Esuli and F. Sebastiani, “SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining,” in *LREC*, 2006, pp. 417–422.

[2] S. Baccianella, A. Esuli, and F. Sebastiani, “Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining,” in *LREC*, 2010, pp. 2200–2204.

[3] A. Irvine and C. Callison-Burch, “A comprehensive analysis of bilingual lexicon induction,” *Computational Linguistics*, vol. 43, no. 2, pp. 273–310, 2017.

[4] J. Huang, Q. Qiu, and K. Church, “Hubless Nearest Neighbor Search for Bilingual Lexicon Induction,” in *ACL*, 2019, pp. 4072–4080.

[5] J. Steinberger, P. Lenkova *et al.*, “Creating Sentiment Dictionaries via Triangulation,” in *WASSA*, 2011, pp. 28–36.

[6] M. Zhao and H. Schütze, “A Multilingual BPE Embedding Space for Universal Sentiment Lexicon Induction,” in *ACL*, 2019, pp. 3506–3517.

[7] P. Dufter, M. Zhao *et al.*, “Embedding Learning Through Multilingual Concept Induction,” in *ACL*, 2018, pp. 1520–1530.

[8] T. Nasukawa and J. Yi, “Sentiment analysis: Capturing favorability using natural language processing,” in *Proceedings of the second international conference on Knowledge capture*, 2003, pp. 70–77.

[9] J. Nivre, M.-C. d. Marneffe *et al.*, “Universal Dependencies v1: A multilingual treebank collection,” in *LREC*, 2016.

[10] A. Conneau, G. Lample *et al.*, “Word Translation Without Parallel Data,” *arXiv preprint arXiv:1710.04087*, 2017.

[11] S. M. Mohammad, B. J. Dorr *et al.*, “Computing Lexical Contrast,” *Computational Linguistics*, vol. 39, no. 3, pp. 555–590, 2013.

[12] D. Tang, F. Wei *et al.*, “Sentiment Embeddings with Applications to Sentiment Analysis,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 2, pp. 496–509, 2016.

[13] Y. Chen and S. Skiena, “Building sentiment lexicons for all major languages,” in *ACL*, vol. 2, 2014, pp. 383–389.

[14] M. Pontiki, D. Galanis *et al.*, “SemEval-2016 Task 5: Aspect Based Sentiment Analysis,” in *SemEval*, 2016, pp. 19–30.

[15] J. Steinberger, T. Brychcín, and M. Konkol, “Aspect-Level Sentiment Analysis in Czech,” in *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. ACL, 2014, pp. 24–30.

[16] J. Ruppenhofer, R. Klinger *et al.*, “IGGSA Shared Tasks on German Sentiment Analysis,” in *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, 2014, pp. 164 – 173.

[17] A. Amram, A. Ben David, and R. Tsarfaty, “Representations and Architectures in Neural Sentiment Analysis for Morphologically Rich Languages: A Case Study from Modern Hebrew,” in *ICCL*, 2018, pp. 2242–2252.

[18] S. Gojali and M. L. Khodra, “Aspect based sentiment analysis for review rating prediction,” in *ICAICTAS*, 2016, pp. 1–6.

[19] P. Basile, D. Croce *et al.*, “Overview of the EVALITA 2018 aspect-based sentiment analysis task (ABSITA),” in *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speechtools for Italian*, 2018.

[20] C. Hashimoto, S. Kurohashi *et al.*, “Construction of a Blog Corpus with Syntactic, Anaphoric, and Sentiment Annotations,” *Journal of Natural Language Processing*, vol. 18, pp. 175–201, 2011.

[21] A. L. Maas, R. E. Daly *et al.*, “Learning Word Vectors for Sentiment Analysis,” in *ACL*, 2011, pp. 142–150.

[22] C. Freitas, E. Motta *et al.*, “Sparkling Vampire...lol! Annotating opinions in a book review corpus,” in *New Language Technologies and Linguistic Research: A two-Way Road*, 2014, pp. 128–146.

[23] 浅原正幸, 金山博 *et al.*, “Universal Dependencies 日本語コーパス,” **自然言語処理** ***26*** **巻** ***1*** **号**, pp. 3–36, 2019.

[24] 金山博, 岩本蘭, “多言語評価表現抽出を通じた Universal Dependencies の検証,” **言語処理学会第** ***26*** **回年次大会予稿集**, 2020.

[25] P. Qi, T. Dozat *et al.*, “Universal Dependency Parsing from Scratch,” in *CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 2018, pp. 160–170.

[26] 金山博, 那須川哲哉, 渡辺日出雄, “木構造変換を利用した評判分析手法,” **人工知能学会論文誌** ***26*** **巻** ***1*** **号**, pp. 273–283, 2011.