DEIM2020 P2-19(day2 p92)

訂正投稿を用いたフェイクニュース収集システムの開発

村山　太一† 　若宮　翔子† 　荒牧　英治†

† 奈良先端科学技術大学院大学
〒 630—0192 奈良県生駒市高山町 8916-5
E-mail: †{murayama.taichi.mk1,wakamiya,aramaki}@is.naist.jp

あらまし　SNS の普及によって多くの人が手軽にニュースに触れやすくなった．一方で，意図して誤った情報を流す「フェイクニュース」と呼ばれる現象が社会問題となっている．「フェイクニュース」に対抗するために，Snopes や poltifact といった SNS 上に大量に流布されるニュースに対し検証を行うサイトも多く生み出されている．しかし，これらの検証サイトは人手で判定を行うことから時間やコストを要するという問題がある．さらに，検証サイトがうまく機能している国は少なく，多くの言語ではそのような検証サイトを活用できないといった問題も存在する．これらの問題に対して，本論文では多言語で自動的にフェイクニュースを収集するシステムを構築する．本収集システムは，「フェイクニュースであることを指摘するツイート」を用いて自動的にフェイクニュースの候補を収集し，様々な言語に拡張が可能なフレームワークとなっている．

キーワード　フェイクニュース，Twitter，大規模システム，データセット

1 はじめに

Twitter や Facebook といった Social Networking Service (SNS) は，コミュニケーションや情報交流の場として積極的に用いられており，多種多様なニュースが溢れている．Pew Research Center の調査によると，アメリカの 20 歳以上の 68% が SNS からニュースを取得している [1]．SNS から容易に多くのニュースを受け取れるようになった，その一方で，ニュースの中には「フェイクニュース」と呼ばれる誤った情報もあり，多く共有されてしまっている．フェイクニュースの広い定義は，[2] において「ニュースの背後に存在する意図などに関わらずメディアなどを通して拡散される記事などのこと」とされている．

フェイクニュースは経済的・政治的理由から広められている場合もあり，社会問題の一つとなっている．例えば，2016 年の米大統領選では，真実とは異なる各候補者にとって有利となるニュースが SNS 上で 3700 万回以上も共有され，選挙結果に大きな影響を与えたといわれている [3], [4]．選挙以外にも，災害 [7], [8]，乱射事件 [5] や株価 [6] などにもフェイクニュースの影響が見られるため，喫緊の課題として対処することが社会的に求められている．

最近では，専門家がニュースの真偽を検証するために，Snopes.com や Politifact.com といったファクトチェッキングサイトが利用されている．同様に，SNS 上のフェイクニュースをトラッキングするツール [9], [10] なども開発されている．このような既存のトラッキングツールでは，ファクトチェッキングサイトによって報告やラベル付けされたニュースを対象に収集を行う．

既存のトラッキングツールはフェイクニュース収集において重要であるが，2 つの問題点が挙げられる．1 点目は，これらのツールの元となっているファクトチェッキングサイトでの検証には多大な時間や労働力を要する点である．SNS 上に様々なフェイクニュースが急速に広まっていることからも，早期の検知は重要である．2 点目に，ファクトチェッキングサイトは主に EU やアメリカといった一部の国々でしか利用できない点である．そのため，主要なファクトチェッキングサイトを持たない英語圏外の国では，既存のトラッキングツールをそのまま用いることは困難である．

これらの問題を解決するために，本論文では人によるアノテーションやチェッキングサイトを必要としないフェイクニュースのトラッキングシステムを提案する．本システムは**「これはフェイクニュースです」**といった SNS ユーザーのコメントが，無料で迅速な信号となるという考えに基づいて構築される．提案システムの特徴は以下の通りである．

- 本システムは，専門家やファクトチェッキングサイトによる判定を用いることなく，SNS ユーザーによる投稿を用いることでフェイクである可能性が高いニュースを収集する．
- 本システムは Twitter を用いて英語と日本語の 2 つの言語で動作する．ルールベース（教師なし）の手法で構成されていることから，様々な言語に容易に拡張が可能である．将来，大規模な多言語フェイクニュースデータセットを公開する予定である．
- 本システムは現在拡散されているフェイクニュースの内容を一般ユーザ向けに提示する．

2 関連研究

2.1 ファクトチェッキングサイト

フェイクニュースに対抗するため様々なファクトチェッキングサイトや組織が設立された．以下に代表的な例を示す．Politi-

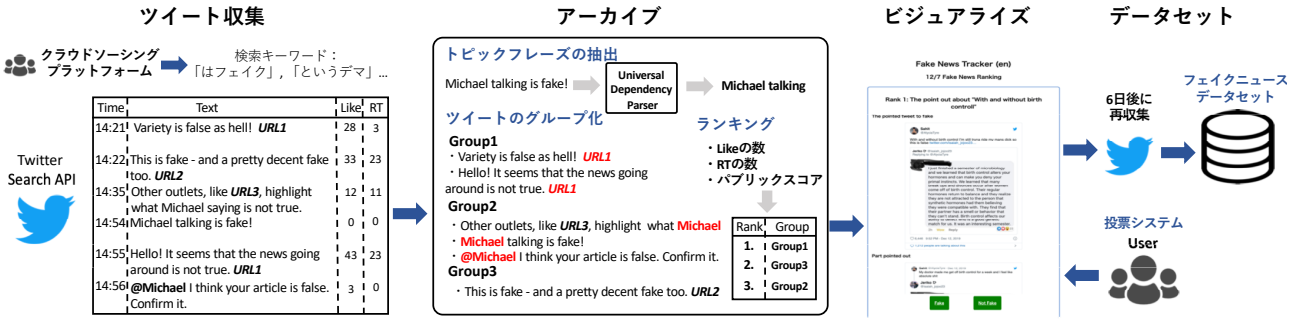

図 1: 提案システムのフレームワーク：システムはツイート収集，アーカイブ，ビジュアライズ，データセットの 4 つの部分で構成されている．ツイート収集では，「フェイク」であると指摘するツイートを収集し，アーカイブで収集したツイートを整理し，ビジュアライズではランク付けされた結果を提示する．データセットでは，将来のフェイクニュースデータセットの構築に向けて再びツイートの収集を行う．

Fact [1]は主に米国の政治ニュースや政治家の発言を対象とした独立で無党派のオンラインファクトチェッキングサイトである．Snopes [2]は初期のチェッキングサイトの 1 つで，政治や社会的イベントや話題となっている事柄を取り扱う．FactCheck [3]は非営利の団体で，テレビ広告や政治家の討論，スピーチといった様々な事柄を対象に検証を行う．Gossipcop [4]は雑誌や Web ニュースに掲載された米国のエンタメニュースを主に対象として検証を行う．これらのチェッキングサイトの信頼度は高いが，検証には時間がかかり，急増しているフェイクニュースに対してスケーラビリティがそこまで高くないという欠点を持つ．

2.2 トラッキングツール

SNS で大量のフェイクニュースが拡散されていることから，そのニュースがフェイクかどうか確認したり，フェイクニュースの性質を調べたりするためにも追跡することは重要である．こういった要求を満たすために，いくつかのトラッキングツールが提供されている．Hoaxy [9] はファクトチェッキング情報やそれに関するフェイクニュースを収集・追跡するためのフレームワークである．Hoaxy を用いることで，興味のあるトピックを検索したり，そのトピックの拡散状況を視覚的に確認することができる．FakeNewsTracker [10] は SNS 上でのフェイクニュースを収集・検知するシステムの 1 つである．これらのトラッキングシステムは，ファクトチェッキングサイトをソースとしてフェイクニュースの収集を行う．NewsVerify [12] はユーザーの入力からニュースの追跡を行い，そのニュースの信頼度を測定するリアルタイムのシステムである． [13] は特定のフレーズからツイートを検索するという点で我々のシステムに近い．「本当？」などの質問のフレーズから噂を収集するが，我々のシステムは「これはフェイクです」といった修正に関するフレーズを用いてフェイクニュースの収集を行う．さらに，我々のシステムは多言語に対して拡張可能なフレームワークとなっている．

1 : https://www.politifact.com/
2 : https://www.snopes.com/
3 : https://www.factcheck.org/
4 : https://www.gossipcop.com/

2.3 トラッキングツールによるデータセット

トラッキングシステムを用いて，フェイクニュース検出などの研究目的で様々なデータセットが作成されている．Hoaxy dataset [14] は，Hoaxy によって収集されたデータセットで，ファクトチェッキングサイトや検証で明らかになったフェイク記事の URL を含んだツイートやリツイートで構成されている．FakeNewsNet [11] は，FakeNewsTracker によって収集されたデータセットであり，対象となった記事の内容とそれについて言及された SNS の投稿やユーザ情報で構成されている．

3 フレームワーク

本章では，最初に提案システムの全体像を示す．次に，提案システムを構成している各部分の詳細を述べる．

3.1 全 体 像

提案システムはツイート収集，アーカイブ，ビジュアライズ，データセット作成の 4 つの部分で構成されている．全体像を図 1 に示す．ツイート収集では「**フェイクである**」と指摘するツイートを収集し，アーカイブでは収集したデータを整理し，視覚化のためにランク付けを行う．ビジュアライズでは注目度の高い順に収集されたツイートを提示し，そのツイートが本当にフェイクニュースであるかどうかをユーザが投票できるページを生成する．データセットでは，フェイクニュースに関する多言語の大規模なデータセットを作成するために，アーカイブで取得した URL やキーワードを用いてツイートの再収集を行う．

3.2 ツイート収集

フェイクであることを指摘しているツイート（指摘ツイート）を収集するために，そのようなツイートに頻繁に含まれている言語パターン（フェイクパターン）を見つける必要がある．有用なフェイクパターンを発見するために，クラウドソーシングを活用する．まず，クラウドソーシングプラットフォームで「Twitter などの SNS で誤った情報 (フェイクニュース) を見つけたとき，どのように修正するか」といった質問に対する回答を収集した．英語のパターンの発見には Amazon Mechanical Turk [5]を，日本語のパターンの発見には Yahoo!ク

5 : https://www.mturk.com/

表 1: 日本語と英語で収集のために選択されたパターン

英語	(isn't|is not) true is (completely) (false|fake) Don ' t believe everything spreading (false|fake) #fakenews
日本語	は (デマ | フェイク) (デマ | フェイク | フェイクニュース) です (フェイク | 間違い | デマ) である というデマ (信じ | 拡散し) ない

ラウドソーシング [6]を用いて，それぞれ 1000 件ずつ回答を収集した．次に，クローリングや検索において有用なフェイクパターンを獲得するため，収集した回答を形態素ごとに分割して，uni/bi/tri/4-gram ごとに抽出し，高頻出なパターンを選択した．これらの高頻出パターンのうち，特定のニュースから独立しており，指摘ツイートを収集したときに高い recall 率を達成するパターンを人手で選択した．ツイート収集に用いるフェイクパターンを表 1 に示す．これらのフェイクパターンを収集キーワードとして，Twitter Search API [7]を用いてツイートのクローリングを行う．クローリングは常に実行され，キーワードを含むツイートがデータベースに保存されるようになっている．

3.3 アーカイブ

現在注目されているフェイクニュースを簡単に確認できるように，収集したツイートを整理する．この部分は，トピックフレーズの抽出，ツイートのグループ化，そしてランキングの 3 つのステップで構成される．これらのステップを収集されたツイート全てに対し行うのは時間がかかり非効率である．そのため，以下の処理は 3 回以上リツイートされたツイートのみに対して毎日 1 回実行される．

3.3.1 トピックフレーズの抽出

このステップでは，指摘ツイートで言及されているトピックフレーズを抽出する．例えば，*"Michael talking is fake!"* というツイートから *"Michael talking"* をトピックフレーズとして抽出する．抽出されたトピックフレーズは次のステップのツイートのグループ化やビジュアライズの際の見出しとしても用いられる．

本システムでは複数言語への拡張を想定し，機械学習ベースではなくルールベースによる抽出を実行する．トピックフレーズ抽出のために，様々な言語に対して一貫した構文構造とタグセットなどが開発された Universal Dependencies (UD) [15] による解析結果を用いる．UD を用いることで様々な言語に対して同様のルールを少しの調整で適用可能である．各言語のツリーバンクは universaldependencies.org [8]から取得し，各ツイートに UD 分析を適用する．得られた構造結果から，人手で決定した抽出ルールを用いてトピックフレーズを抽出する．抽出ステップを以下に示す．

（1） UD 解析を用いてフェイクパターン（表 1）を含むツイート内の文章を解析する．本解析では，フェイクパターンにマッチする部分を「フェイクパート」と呼称する．

（2） そのフェイクパートに対する依存構造が，主格·目的格·同格·名詞句（"nsubj"，"nsubjpass"，"dobj"，"iobj"，"csub"，"appos"）などで掛かっており，かつ，フェイクパートより前に出現する部分をトピックフレーズとして抽出する．ただし，指示語（これ，それ）の場合は抽出部分なしとする．

（3） (2) でトピックフレーズが存在しない場合，現在のフェイクパートが掛かる語をフェイクパートと設定し，(3) を実行する．

（4） (3) で掛かる語が存在しない場合，つまりフェイクパートが root である場合，ツイート内の探索する文を変更する（日本語の場合は 1 文前，英語の場合は 1 文後）．変更された文の root をフェイクパートとして，(2) を実行する．

3.3.2 ツイートのグループ化

このステップでは，言及されているイベントが同様のものであれば同じグループとして設定する．Twitter では毎日多種多様なイベントに関する言及が飛び交っていることからも，機械学習を適用してグループ化することは難しい．よって，3.3.1 節で抽出したトピックフレーズやツイートに付与された URL などを用いて単純で頑健なルールベースの手法でグループ化を行う．グループ化のためのルールは以下の通りである．

（1） 同じ URL について言及しているツイートを同一グループとして設定．

（2） 同じツイートに対するリプライを同一グループとして設定．

（3） 抽出されたトピックフレーズ同士の距離を Word Mover's distance (WMD) [16] を用いて測定し，距離が閾値 τ 以下のものを同一グループとして設定．

WMD の計算のために [17] で提供されている単語ベクトルを用いた．また，閾値 τ はどちらの言語についても 0.25 と設定した．

3.3.3 ランキング

3.3.2 節でグループ化されたイベントを，注目度の高さに基づきランク付けする．今回用いるランキングの手法は，複数の特徴を用いて各グループのランク付けを行うという [18] で用いられた教師なし手法を参考に設定する．具体的には，注目度を示していると考えられるいくつかの特徴をランク付けし，それらの平均ランクを最終的なランキングとして算出する．本シス

6 : https://crowdsourcing.yahoo.co.jp/

7 : https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets

8 : https://universaldependencies.org/

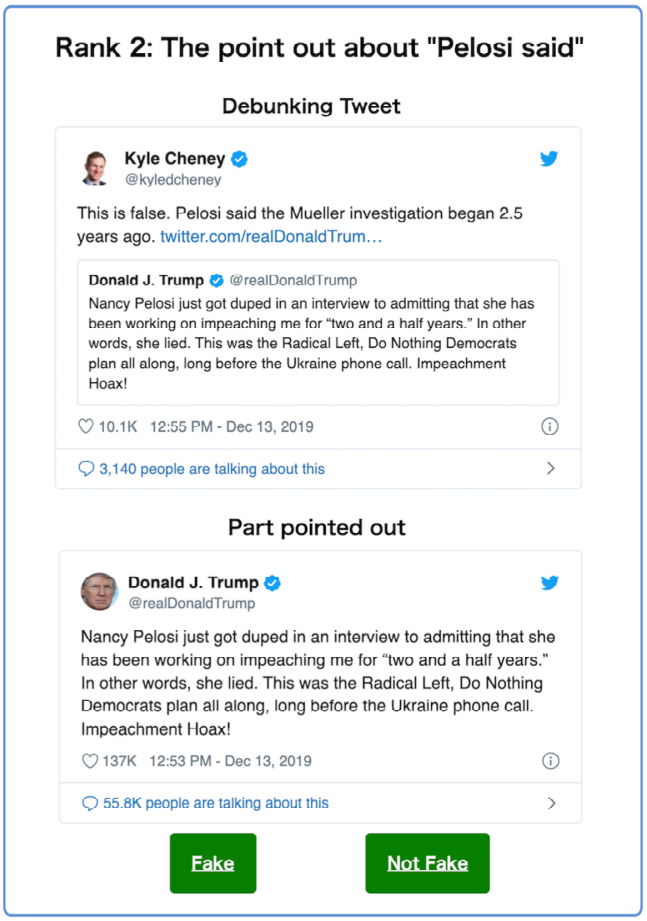

図 2: 提案システムの視覚化の例．各イベントは「見出し」,「指摘ツイート」,「指摘されている部分」,「投票システム」の 4 つの部分から構成される．

テムでは **Like 数**，**RT 数**，**パブリックスコア**の 3 つの特徴を用いる．パブリックスコアは，リツイート (RT) を行ったユーザの中に含まれるフォロワーの割合を求めた数値である．この値が小さければ小さいほどフォロワー以外のユーザに情報が広がっており，注目度が高いといえる．一方，Like 数と RT 数については大きいほど注目度が高いとし，これらの各特徴量のランク付けを行う．

3.4 ビジュアライズ

現在注目されているイベントをランク付けで表示する．表示例を図 2 に示す．

本システムでは，注目度の高い上位 10 件のニュースを毎日更新して掲載する．各ニュースは「見出し」「指摘ツイート」「指摘されている部分」「投票システム」の 4 つの部分で構成される.「見出し」では，3.3.1 節の処理で抽出されたトピックフレーズを用いて，内容を簡潔に説明する.「指摘ツイート」では，本システムで収集したツイートを,「指摘されている部分」では収集したツイートに含まれる URL や引用先，リプライ先の情報を，それらが無ければ「見出し」を記載する．本システムでは，ファクトチェッキングサイトを用いずに，フェイクである可能性が高いニュースイベントを収集している．そのため，イベントが実際にフェイクであるか否かが不明である．そこで，

表 2: 収集されたツイートの統計量（1 日の平均を記載）

	英語	日本語
収集される平均ツイート数	9039	7901
収集される平均イベント数	455	143
上位 10 件の平均 RT 数	1549	810
上位 10 件の平均 Like 数	5027	1889
上位 10 件の平均認証バッジアカウント数	2.11	0.20

「投票システム」では，各イベントに対してユーザがフェイクかどうかを投票できる機能を実装し，ユーザの投票結果に基づくラベル付けを行う．このように，現在注目されているフェイクニュースの概要をわかりやすく表示する．

3.5 データセット

将来，本システムで収集されるツイートをもとに，フェイクニュースデータセットを構築する予定である．また，英語や日本語以外の言語にも拡張し，大規模な多言語データセットを構築する予定である．

各言語の網羅的なデータセット作成のために，本システムで視覚化される 10 件/日のイベントを対象に，言及される URL や 3.3.1 節で取得したトピックフレーズをキーワードと設定し，再びクローリングを行う．再び収集されたツイートと，3.4 節で投票されたラベルを元に多言語データセットを構築し公開する予定である．

4 議　　論

本章では，提案システムで収集されるツイートの傾向と有効性について検証する．

4.1 統　計　量

2019 年 11 月 14 日から 2019 年 12 月 13 日までに英語・日本語の 2 カ国語で収集されたツイート数やイベント数などの統計を表 2 に示す．

表 1 で設定されたキーワードで収集されたツイート数は日本語と英語で大きな差が見られなかった．しかし，グループ化によって作成されたイベント数は英語が日本語の 3 倍程度となった．このことは，日本語で収集されたツイートの多くが，リツイートによるものだったことが要因だと考えられる．また，取得された上位 10 件の英語のイベントが日本のイベントと比較して，RT 数，Like 数ともに多かった．これは認証バッジを持った影響力のあるアカウントによる訂正が，日本語よりも英語のほうが圧倒的に多かったためであると考えられる．

4.2 収集システムの有効性

提案システムが適切に動作しているかを確認するために，ランク付けされたニュースがフェイクニュースかどうかの確認を行う．2019 年 12 月 7 日から 12 月 13 日にランク付けされた指摘ツイート 124 件（英語 62 件，日本語 62 件）を対象に，以下の観点でアノテーションする．

(a) 収集された指摘ツイートはフェイクであることを指摘するツイートであるか？

(b) 指摘ツイートが指摘している内容は，実際にフェイクやデマであるか？

1 章に示した「フェイクニュース」の定義に従いコードブックを作成し，2 名のアノテーターがアノテーションを行った．この結果，2 名の Cohen's Kappa score は 0.73 で一定の合意をとれたことを確認した．なお，2 人のアノテーターが同意しなかったニュースについては第 3 の評価者（著者のうち 1 名）がラベルを付与した．

(a) の観点について，評価結果から，日本語・英語ともに 6 割以上（日本語では 66%，英語では 69%）の指摘ツイートが，適切にフェイクであることを指摘する形式をとることが分かった．このことは，3.1 節で選択したフェイクパターンがある程度適切であったことを示唆する．(b) の指摘された内容がフェイクであるかどうかについて，日本語・英語とともにランキングで提示したニュースのうち 6 割以上（日本語では 68%，英語では 65%）が実際にフェイクであるという結果が得られた．システムのアーキテクチャは言語に依らず同一のものを用いているが，適切にフェイクニュースの可能性が高いものを取り上げている．この結果は，フェイクニュース収集システムとして実用可能な水準であることを示している．

5 おわりに

本論文では，ニュースを訂正する投稿に着目し，新たなフェイクニュース収集システムの提案・検証を行った．提案システムは現在日本語と英語で稼働している．容易に多言語に拡張可能という特徴から，今後は他の言語にも拡張し大規模なデータセットを公開する予定である．多言語のフェイクニュースを収集することで，これまであまり行われてこなかった言語や国ごとのフェイクニュースの比較がより容易になると考えられる．

謝　辞

本研究の一部は，JSPS 科研費 19K20279, 19H04221，および厚生労働省科学研究費補助金（課題番号： H30-新興行政-指定-004）の支援を受けたものです．

文　献

[1] Shearer, E and Matsa, K, "News use across social media platforms 2018", Pew Research Center, Journalism and Media, 2018.

[2] Sharma, Karishma, et al, "Combating fake news: A survey on identification and mitigation techniques.", ACM Transactions on Intelligent Systems and Technology (TIST), 10.3, 21, 2019.

[3] Budak, Ceren, "What happened? The Spread of Fake News Publisher Content During the 2016 US Presidential Election.", The World Wide Web Conference, pp.139–150 , 2019.

[4] Bovet, Alexandre, and Hernán A. Makse, "Influence of fake news in Twitter during the 2016 US presidential election.", Nature communications, 10.1, 7, 2019.

[5] Starbird, Kate, "Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter.", In Proc. of International AAAI Conference on Web and Social Media, pp.230–239, 2017.

[6] Fortune, Fake Bloomberg news report drives Twitter stock up 8%., Available at `http://fortune.com/2015/07/14/fake-twitter-bloomberg-report/` 2015.

[7] Mendoza, Marcelo, Barbara Poblete, and Carlos Castillo, "Twitter under crisis: Can we trust what we RT?", In Proc. of the first workshop on social media analytics, pp.71–79, 2010.

[8] Takayasu, Misako, et al., "Rumor diffusion and convergence during the 3.11 earthquake: a Twitter case study.", PLoS one, 10.4, e0121443, 2015.

[9] Shao, Chengcheng, et al., "Hoaxy: A platform for tracking online misinformation.", In Proc. of the 25th international conference companion on world wide web., pp.745–750, 2016.

[10] Shu, Kai, Deepak Mahudeswaran, and Huan Liu., "FakeNewsTracker: a tool for fake news collection, detection, and visualization.", Computational and Mathematical Organization Theory, 25.1, pp.60–71, 2019.

[11] Shu, Kai, et al., "FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media.", arXiv preprint, ArXiv:1809.01286, 2018.

[12] Zhou, Xing, et al., "Real-Time News Certification System on Sina Weibo.", In Proc. of the 24th International Conference on World Wide Web., pp.983–988, 2015.

[13] Zhao, Zhe, Paul Resnick, and Qiaozhu Mei., "Enquiring minds: Early detection of rumors in social media from enquiry posts.", In Proc. of the 24th International Conference on World Wide Web., pp.1395–1405, 2015.

[14] Hui, Pik-Mai, et al., "The Hoaxy misinformation and fact-checking diffusion network." In Proc. of the twelfth International AAAI Conference on Web and Social Media., pp.528–530, 2018.

[15] McDonald, Ryan, et al., "Universal dependency annotation for multilingual parsing.", In Proc. of the 51st Annual Meeting of the Association for Computational Linguistics, pp.92–97, 2013.

[16] Kusner, Matt, et al., "From word embeddings to document distances.", In Proc. of the International conference on machine learning., pp.957–956, 2015.

[17] Grave, Edouard, et al., "Learning word vectors for 157 languages.", In Proc. of the International Conference on Language Resources and Evaluation, 2018.

[18] Glavaš, Goran and Štajner, Sanja., "Simplifying lexical simplification: Do we need simplified corpora?", In Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing., pp.63–68, 2015.