卒業論文

Twitterを用いた新聞記事への
自動ソーシャルアノテーションツールに関する研究

平成 26 年 2 月 7 日提出

指導教員　　伊庭斉志 教授
ダヌシカ ボレガラ 講師

電子情報工学科

03-120443　　村上 晋太郎

目 次

1 序論 **1**
1.1 背景 1
1.2 目的 2

2 手法 **3**
2.1 記事と関連するニュース記事の収集 3
2.2 内容語の抽出 4
2.2.1 tf-idf 4
2.3 関連度の計算 5
2.4 言い換えの関係の語の検出 6
2.5 互いに類似する語の検出 6
2.5.1 WordNet を用いた類似語の検出 6
2.5.2 Dekang Lin の手法の利用 7
2.5.3 Latent Space でのモデル化 8
2.6 アラインメント 9
2.7 形態素解析 10

3 実験 **10**
3.1 実験概要 10
3.2 実験条件 11
3.3 実験結果 11
3.3.1 内容語の抽出 11
3.3.2 アラインメント 12

4 考察 **12**
4.1 内容語の抽出 12
4.2 アラインメント 13

5 今後の課題 **13**
5.1 データ収集の自動化 13
5.2 内容語の抽出 13
5.3 言い換え語・類似語の検出 13
5.4 ユーザー評価の実施 14

6 結論 **14**

1 序論

1.1 背景

2000年代後半から、Facebook, Twitter, Google+などのソーシャルネットワーキングサービスが盛んに利用されるようになった。2012年には、Facebookのアクティブユーザ数は10億人[1]を突破し、Twitterでもアクティブユーザー数は1.4億人[2]となっている。Facebookでは一日当たり0.5ペタバイトのデータが増加しており、Twitterへの投稿数は一日当たり3.4億投稿となっている。このように、ソーシャルネットワーキングサービスは利用者数の面でも、データ量の面でも莫大な資源となっており、様々な利用が期待される。

一方、近年インターネット上で閲覧することが出来るニュース記事が増加している。日本では、朝日新聞社[3]、読売新聞社[4]、毎日新聞[5]等の大手新聞者がインターネット上でニュース記事を配信している。海外でも、Washington Post[6] Las Angels Times[7]等、大手新聞社が同様のサービスを提供している。

ソーシャルネットワーキングサービスの書き込みの中には、このようなインターネット上のニュース記事に言及しているものが多数存在する。このような書き込みには、一般のインターネットユーザーの個人的な意見、見解が記されている。このようなインターネット上のニュース記事と、ソーシャルネットワーキングサービスへの書き込みを結びつけることで、報道機関のプロの書き手とソーシャルネットワーキングサービス上で活動する一般社会の書き手を結びつけることができるようになる。このようなツールが実現すれば、報道機関のプロの書き手は自分の書いた記事が一般社会にどのような影響を与えているのかについての情報を、より簡単に得る事が出来るようになる。一方、一般の書き手にとっては、ニュース記事に情報が付与されるようになり専門的な内容の記事を読みやすくなったり、ニュース記事からより有用な情報を引き出せるようになる。

また、ニュース記事を学術論文に、ソーシャルネットワーキングサービス上の投稿を学術論文に対するコメントに置き換えた場合も、同じ手法を用いることができ、その場合にも同じ効果が期待できるという点で、応用性も高いと考える。

以上のような背景から、本研究ではソーシャルネットワーキング上の情報とインターネット上のニュースを自然言語処理の手法を用いて結びつけることを目標とする。

[1] http://ja.wikipedia.org/wiki/Facebook
[2] http://www.724685.com/twitter/tw13050310.htm
[3] http://www.asahi.com
[4] http://www.yomiuri.co.jp
[5] http://mainichi.jp
[6] http://www.washingtonpost.com
[7] http://articles.latimes.com

1.2 目的

本研究では、ソーシャルネットワーキングサービス「Twitter」上の投稿をインターネット上のニュース記事に自動付与 (アノテーション) するツールを作成することを目的とする。本研究に置けるツールでは、ニュース記事単位を Twitter 上の投稿の付与の対象とするのではなく、ニュース記事の中の文単位を付与の対象とする (図 1)。結果、ユーザはニュース中のどの部分に対して、Twitter 上でどのような反応が起こっているのかが分かるようになる。このようなツールが実現すれば、ニュース記者にとってはは自分の書いた記事についてのユーザーの意見が簡単に見れるようになり、フィードバックをすぐに得る事ができるようになるという点で有用である。一方、読者にとっては、難しい内容のニュース記事でも他のユーザーによる情報付与があることにより、理解の助けとすることができる。Twitter 上の投稿を Web 上のニュース自体に結びつけるタスクは、先行研究で既に取り扱われているが、Twitter 上の投稿を Web 上のニュースの中の断片に結びつけるというタスクはまだ取り扱われていない。その点が、本研究の新規性である。

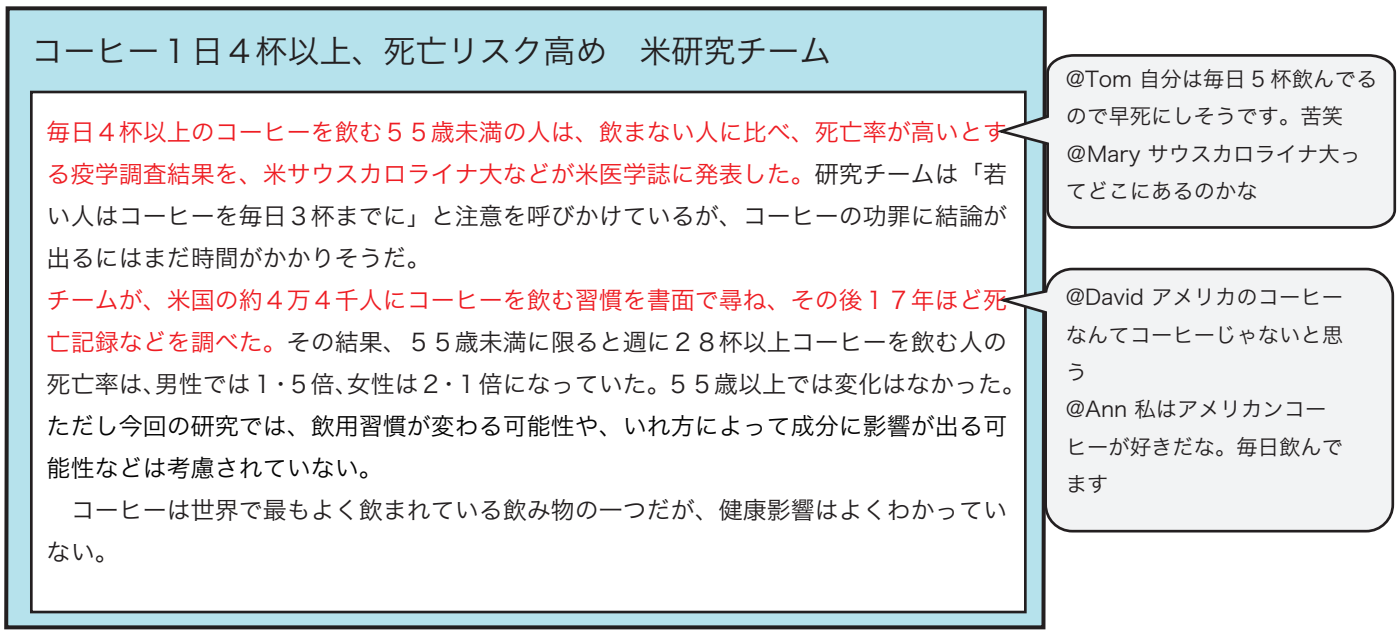

図 1: Twitter を用いたソーシャルアノテーションツール

本研究では Twitter とインターネット上のニュース記事を対象にツールの構築を進めていくが、その本質は「大きな文章中のセンテンスの集合と、小さな文章の集合のアラインメント」である。そのため、「背景」でも述べたように、本研究で得られた成果を、Twitter やインターネット上のニュース記事に限らずに応用することが可能である。例えば、ニュースを学術論文に置き換え、Twitter 上の投稿をその論文に対するレビューに置き換えることで、その学術論文のどの部分にどのような反応が起きているかを知る事が出来るツールを構築可能である。

2　手法

本研究では、ニュース上の記事の断片に対して、ソーシャルネットワーキングサービス「Twitter」上の投稿から関連度の高いものを結びつけるツールを作成する。この章では、そこで用いる手法について説明する。

本研究のタスクは、以下のようなステップに分けることができる。

1. 記事と関連するニュース記事の収集
   Twitter 上の投稿の中から、対象のニュース記事と関連性が高いと思われるものを抽出して収集する。

2. 内容語の抽出
   Twitter 上の投稿の文中から、Twitter 上の投稿とニュース断片の関連度の計算にほとんど影響しないと判断できる単語を除去する。

3. 関連度の計算
   Twitter 上の投稿と、ニュースの断片との関連度を計算して数値化する。

4. ニュース断片と Twitter 上の投稿の結びつけ
   計算された Twitter 上の投稿とニュース断片の関連度をもとに、ニュース断片への Twitter 上の投稿の結びつけを行う。

以下で、上記のそれぞれのステップについて詳しく説明をする。

2.1　記事と関連するニュース記事の収集

記事と関連するニュースの収集に関しては、同じタスクを取り扱った先行研究がすでにあるため本研究では深く取り扱わない。以下、本研究における暫定的な手法について説明する。

Twitter 上においては、あるニュースに関する投稿をする際に、そのニュースのタイトルを併記するという慣習がある。そこで、本研究では、対象のニュースのタイトルを含む Twitter 上の投稿を、その記事と関連するニュースとして自動的に収集するようにした。しかし、Twitter 上の投稿の中には、

- タイトルを併記せずにそのニュースに言及しているもの
- 誤植などで不完全な形でタイトルを併記してそのニュースに言及しているもの
- そのニュースに言及はしていないが、偶然タイトルを含んでいるもの

が存在する。また、他のソーシャルネットワーキングサービス上で同様の慣習が根付いているとは限らないため、応用の範囲も狭められる。今後先行研究をもとに記事の収集を別の手法に置き換えることが望ましいと考える。これについては今後の課題としたい。

2.2 内容語の抽出

自然言語の文の特性を調べる際に、bag-of-words[1] という考え方がる。bag-of-words では、ある文の特性を解析する際に、その文にどのような単語がどのくらい含まれているのかを計算して、それをもとに文の特性を解析し、クラスタリング等のタスクを行う。本研究でも、web ニュースの断片に含まれる単語と、Tweet に含まれる単語の集合を比較して、web ニュースの断片と Tweet の関連度を計算する。

一般に、自然言語の文中には、その文章の持つ素性を計算する際にほとんど情報をもたないような単語が多く含まれる。副詞や格助詞などがそうである。反対に、その文章の持つ素性を計算する際に有用な情報を与える単語を内容語と定義する。非内容語である単語をあらかじめ除去し、内容語を抽出しておくことで、関連度の計算の精度を高めることが出来ると考えられる。関連度の計算の前処理として、この内容語の抽出を行う。

内容語の抽出には、二通りの方法が考えられる。一つ目は、形態素解析ツールを使用し、単語の品詞を分析した上で、副詞や格助詞などの非内容語であると考えられる単語を除去する方法である。二つ目は、統計的な手法を用いて、その単語のその文章における重要度を計算し、一定の重要度以下の値を持つ単語を非内用語として除去する方法である。

一つ目の方法は比較的簡単に実現でき、一定の精度も期待できるが、

- 形態素解析ツールに依存するため他言語に拡張できない
- ソーシャルネットワーキングサービス特有の、ネットスラングなどの語彙に対応できない

といった問題点が存在する。そこで、本研究では二つ目の、統計的手法を採用する。

2.2.1 tf-idf

本研究では、統計学的手法として tf-idf[2] を採用する。tf-idf は、文書中である単語が、その文章をどの程度特徴づけているかの重み度合いである。tf-ifd は、以下の式によって定義される。

$$\mathrm{tfidf_{w,d}} = \mathrm{tf_{w,d}} \cdot \mathrm{idf_w} \tag{1}$$

$$\mathrm{tf_{w,d}} = \frac{\mathrm{n_{w,d}}}{\sum_{\mathrm{k}} \mathrm{n_{w,d}}} \tag{2}$$

$$\mathrm{idf_w} = \log \frac{|\mathrm{D}|}{|\{\mathrm{d}|\mathrm{t_w} \in \mathrm{d}\}|} \tag{3}$$

ここで、$n_{w,d}$ は単語 w の文書 d における出現回数、$|D|$ は総ドキュメント数、$|\{d|t_w \in d\}|$ は単語 w を含む総ドキュメント数である。

この式の意味するところは、「どのような文書にも普遍的に出現するような単語の重みは低くなる」ということと、「ある文書中で頻繁に言及されるような単語があれば、その文書中でのその単語の重みは高くなる」ということである。

本研究では、tf-idf の高い単語は内容語であると判断し、tf-idf の低い単語は内容語では無いと判断する。

本研究では、まず新聞記事を 100 件以上集め、idf の計算を行う。ここで、収集する記事が多ければ多いほど idf の精度は上がる。この ifd をもとに、web ニュースの断片と tweet の関連度を

計算する際に tf-idf を計算し、単語を tf-idf でソートする。そして、その中で閾値以上の単語を内容語とする。この閾値を調整すると、内容語の数と質が変わるので、実験によって調整する。

2.3 関連度の計算

内容語の抽出を行ったのちに、Twitter 上の投稿とニュース断片の関連度の計算を行う。関連度の計算にも、bag-of-words の考え方を採用し、文中に含まれる単語を分析する。

単語をもとにニュース断片と Twitter 上の投稿の関連度を計算する際の指標として、以下の三つが挙げられる。

1. 完全に一致する単語
   web ニュースの断片と Twitter 上の投稿の間に完全に一致する単語がある場合、そのセンテンスと Tweet の関連度は高いと考えられる。図 1 の例では、@Mary というユーザーが投稿した tweet に「サウスカロライナ」という単語が含まれているので、同じ単語を含む一つ目のセンテンスとの関連度が高くなる。

2. 言い換えの単語
   web ニュースの断片と Twitter 上の投稿に言い換えの関係にある単語が含まれている場合、そのセンテンスと Tweet の関連度は高くなると考えられる。図 1 の例では、@David というユーザーが投稿した tweet に「米国」と言い換えの関係にある「アメリカ」という単語が含まれているので、「米国」という単語が含まれる、5 行目から始まるセンテンスとの関連度が高くなる。

3. 互いに類似する単語
   web ニュースの断片と tweet の文に、同じ文脈で使われるような単語が含まれている場合、そのセンテンスと Tweet の関連度は高くなると考えられる。図 1 の例では、@Tom というユーザーが投稿した Tweet に「早死に」という単語が含まれており、これは「死亡率」という単語と同じ文脈で使われる単語であると判断できるので、一行目から始まるセンテンスとの関連度が高くなる。

これらの単語が含まれていたら値が高くなるようにセンテンスと Tweet の関連度を定義する。このようにすることで、ニュース web ニュースの断片と Twitter 上の単語を比較し、のアラインメントをとることができる。

上記の完全に一致する単語、言い換えの単語、互いに類似する単語の中でも、単語によってニュース記事と Twitter 上の投稿の関連度に対する影響は異なると考えられる。そこで、本研究ではこの影響の違いを tf-idf で重み付けすることで対応する。また、「一致する単語」、「言い換えの単語」、「互いに類似する単語」の三つの集団でも、ニュース記事と Twitter 上の投稿の関連度に対する影響は異なると考えられる。これは定数で重み付けをすることで対応する。この定数は、実験で調整する。

以上より、Tweet$t$ とニュース断片 s の関連度を以下の式によって定義する。

$$\mathrm{Rel}(s,t) = \alpha \sum_{\{w|w\in A\}} w * \mathrm{tfidf}_{w,d} + \beta \sum_{\{w|w\in B\}} w * \mathrm{tfidf}_{w,d} + \gamma \sum_{\{w|w\in C\}} w * \mathrm{tfidf}_{w,d} \qquad (4)$$

ここで、A は s と t の中で完全に一致する単語の集合であり、B は s と t の中で言い換えの関係にある単語の集合であり、C は s と t の中で互いに類似の関係にある単語の集合である。同じ単語が複数回出てきた場合も、それは別々の単語として処理する。すなわち、tf-idf の高い単語が s と t の両方に複数回にわたって出現していたら、それだけ関連度 $\mathrm{Rel}(s,t)$ は高くなる。α, β, γ は定数であり、実験によりアラインメントの結果を見ながら調整する。「完全に一致する単語」の方が、「言い換えの単語」や「互いに類似する単語」よりも関連度に大きな影響力を持つと考えられる。また、「言い換えの単語」は「互いに類似する単語」よりも関連度に大きな影響力を持つと考えられる。そのため、$\alpha > \beta > \gamma$ となることが予想される。

上記であげた蜜の集合のうち、A の「完全に一致する単語」は容易に発見することができる。それに対して、B の「言い換えの単語」や、C の「互いに類似する単語」を識別するためには、統計的な処理が必要となる。この手法については後述する。

2.4 言い換えの関係の語の検出

本研究では「言い換えの単語」の検出方法として、WordNet[8]を利用する。WordNet は、英語の概念辞書である。WordNet では、単語が synset という同義語のグループにまとめられている。よって、同一の synset に含まれる単語は言い換えの単語であると判断できる。WordNet には英語版の他にも、独立行政法人情報通信研究機構（NICT）によって日本語版の物が提供されている。また、Python 等のプログラミング言語向けのフロントエンドも提供されており、プログラムから動的にデータ資源にアクセスすることが可能である。

2.5 互いに類似する語の検出

「互いに類似する単語」の検出にはいくつかの方法が考えられるが、本研究では言い換えの関係の語の検出と同じく、WordNet を利用して類似する単語の検出を行う。しかし、WordNet を用いた類似語の検出にはいくつかの問題点が存在するため、考えられる代替案として Dekang Lin の手法 [3]、Latent Space でのモデル化 [4]、Co Clustering[5] を後に紹介する。

2.5.1 WordNet を用いた類似語の検出

WordNet では同義語同士が Synset というグループにまとめられているが、各々の Synset は他の Synset と繋がりを持っている。Synset 同士の繋がりにはいくつかの種類がある。Synset 同士の繋がりの種類は、単語の品詞によって異なる。以下に、名詞の Synset における Synset 間の繋がりの例を示す。

- 上位語
  すべての X が Y の種類の一であるなら Y は X の上位語である。
- 下位語
  すべての Y が X の種類の一であるなら Y は X の下位語である。

[8] http://nlpwww.nict.go.jp/wn-ja/

- 同族後
  X と Y の上位語が同じなら、Y は X の同族語である。
- 全体語
  X が Y の一部であるなら、Y は X の全体語である。
- 部分語
  Y が X の一部であるなら、Y は X の部分語である。

以上のような繋がりを用いて、WordNet 中では Synset 同士が階層構造を形成している。Synset のネットワークの例を図 2 に示す。

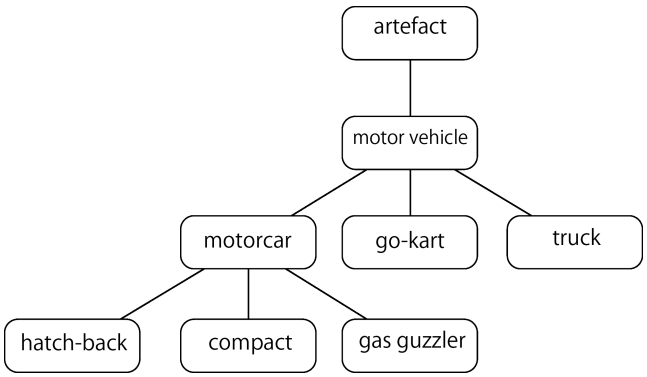

図 2: WordNet における Synset 同士の繋がりの例

図 2 において、artefact は motor vehicle の上位語であり、motorcar, go-kart, truck はともに motor vehicle の下位語であるので、互いに同族語である。

このようなネットワーク上では、意味の近い単語ほど短い経路でたどり着くことができ、意味の遠い単語ほどたどり着くのに長い経路を必要とすると考えることができる。このネットワーク上での距離を測定することで、Synset 間の類似度を測定することが出来る [8]。本研究ではこの性質を利用して単語間の類似度を測定する。そして、類似度が一定の閾値以上の単語同士を互いに類似する単語として集計する。この類似度は実験により調整する。

2.5.2 Dekang Lin の手法の利用

WordNet による類似度の測定は効果的であるが、人手によって構築された意味辞書に依存しているという欠点がある。英語圏には既に完成度が高い WordNet が存在するが、それをもとに作られた日本語圏の WordNet はまだ語彙数が少ない。このように、WordNet を用いた類似語の検出は言語に依存してしまう。また、ソーシャルネットワーキングサービス上ではネットスラング等の辞書上に無い新しい語彙が用いられる場合もあり、その場合も WordNet では対応することが出来ない。

以上のような理由から、本研究のツールにおける類似語の検出には意味辞書に依存せずに、収集したテキストコーパスから学習できるような手法が望ましいと考えられる。

このような「互いに類似する語」の検出方法として、Dekang Lin の手法 [3] の利用が挙げられる。これは、単語同士がどれだけ「似ているのか」を計算するための手法である。Dekang Lin の手法では、単語同士の類似度は以下の式で定義される。

$$\mathrm{sim}(w_1, w_2) = \frac{\sum_{(r,w)\in T(w_1)\cap T(w_2)} (I(w_1, r, w) + I(w_2, r, w))}{\sum_{(r,w)\in T(w_1)} I(w_1, r, w) + \sum_{(r,w)\in T(w_2)} I(w_2, r, w)} \tag{5}$$

ここで、I は相互情報量であり、以下の式で定義される。

$$I(w, r, w') = -\log(P_{MLE}(B)P_{MLE}(A|B)P_{MLE}(C|B)) - (-\log P_{MLE}(A, B, C)) \tag{6}$$

$$= \log \frac{||w, r, w|| \times ||*, r, *||}{||w, r, *|| \times ||*, r, w'||} \tag{7}$$

ここで、$||w, r, w'||$ という記法は、単語 w, w'、単語間の関係 r に対して w が w' の r であるという関係 (w, r, w') が成り立つ箇所の総数である。例えば、"I have a brown dog"という文章では、w を"dog"、w' を"have"、r を"object of"として、$(\mathrm{dog}, \mathrm{objectof}, \mathrm{have})$ という関係が成り立っているので、$||\mathrm{dog}, \mathrm{objectof}, \mathrm{have}|| = 1$ となる。w, w' は $*$(ワイルドカード) で置き換えられることもあり、その場合、$*$ に全ての単語をあてはめた総数が $||w, r, w'||$ の値となる。

P_{MLE} は最尤法に基づいて計算される確率であり、P_{MLE} の引数となっている A, B, C はそれぞれ以下のような事象である。

A**:** 無作為に選出した語が w である

B**:** 無作為に選出した関係が r である

C**:** 無作為に選出した語が w' である

P_{MLE} は以下の式を満たすことが示されている。

$$P_{MLE}(B) = \frac{||*, r, *||}{||*, *, *||} \tag{8}$$

$$P_{MLE}(A|B) = \frac{||w, r, *||}{||*, r, *||} \tag{9}$$

$$P_{MLE}(C|B) = \frac{||*, r, w'||}{||*, r, *||} \tag{10}$$

これらの定義の意味するところは、同じ単語と同じ関係にある二つの単語は、似ている単語であるということである。例えば、「車」と「電車」は共に「乗る」という動詞の目的語になるので、似ている単語と見なすことができる。

本研究では、この定義に従って $\mathrm{sim}(w_1, w_2)$ を計算し、この値が閾値を越えたペアを互いに類似する単語とみなす。この閾値を調整すると、検出される類似語のペアの質が変わるので、実験によって最適な値を求める。

2.5.3 Latent Space でのモデル化

意味辞書に依存せずに「互いに類似する語」のもう一つの検出方法として Latent Space でのセンテンスのモデル化 [4] が挙げられる。この手法では、

$$X_{ij} = \mathrm{tfidf}_{i,j} \tag{11}$$

によって定義される $M \times N$ 行列 X を、$M \times K$ 行列 P, $N \times K$ 行列 Q を用いて

$$X \approx P^T Q \tag{12}$$

と近似的に分解することにより、次元数の低い素性を抽出する手法である。ここで、P, Q はランダムに初期化された上で

$$P_{\cdot,i} = (Q\tilde{W}^{(i)}Q^T + \lambda I)^{-1} Q\tilde{W}^{(i)}X_{i,\cdot}^T \tag{13}$$
$$Q_{\cdot,i} = (P\tilde{W}^{(j)}P^T + \lambda I)^{-1} P\tilde{W}^{(i)}X_{\cdot,j}^T \tag{14}$$

によって段階的に求められる。$W^{(i)}$ は、

$$W_{i,j} = \begin{cases} 1 & \text{if } X_i j \neq 0 \\ w_m & \text{if } X_i j = 0 \end{cases}$$

によって定義される W の i 行目を成分とする対角行列で、

$$W^{(i)} = \mathrm{diag}(W_{\cdot,i}) \tag{15}$$

と表される。

行列 P の各列には、X に現れていた、各単語の「どの文書でどのくらい重要になるか」という特徴量が低次元に圧縮されて格納されている。これは、単語の情報が抽象化されて入っていると言える。この P の中で、コサイン類似度の高い列同士のペアを見つければ、その列に割り当てられている単語同士は互いに類似する単語と言えるようになる。本研究では、コサイン類似度が閾値を越えたペアを類似語とみなす。この閾値を調整すると、検出される類似語のペアの質が変わるので、実験によって最適な値を求める必要がある。

2.6 アラインメント

上記の手法で検出された「完全に一致する単語」「言い換えの単語」「互いに類似する単語」をもとに、ニュース web ニュースの断片と tweet の関連度を計算し、ニュース web ニュースの断片と tweet がどのように対応しているのかを算出し、アラインメントをとる。ニュース記事 d 中のセンテンス s と tweet$t$ の関連度は、以下の式によって定義する。

$$\mathrm{Rel}(s,t) = \alpha \sum_{\{w|w\in A\}} w * \mathrm{tfidf}_{w,d} + \beta \sum_{\{w|w\in B\}} w * \mathrm{tfidf}_{w,d} + \gamma \sum_{\{w|w\in C\}} w * \mathrm{tfidf}_{w,d} \tag{16}$$

ここで、A は s と t の中で完全に一致する単語の集合であり、B は s と t の中で言い換えの関係にある単語の集合であり、C は s と t の中で互いに類似の関係にある単語の集合である。同じ単語が複数回出てきた場合も、それは別々の単語として処理する。すなわち、tf-idf の高い単語が s と t の両方に複数回にわたって出現していたら、それだけ関連度 $\mathrm{Rel}(s,t)$ は高くなる。α, β, γ は定数であり、実験によりアラインメントの結果を見ながら調整する。「完全に一致する単語」の方が、「言い換えの単語」や「互いに類似する単語」よりも関連度に大きな影響力を持つと考えられる。また、「言い換えの単語」は「互いに類似する単語」よりも関連度に大きな影響力を持つと考えられる。そのため、$\alpha > \beta > \gamma$ となることが予想される。

計算された関連度 $\mathrm{Rel}(s,t)$ がある閾値よりも大きければ、s と t は互いに対応する関係にあると結論づけることができる。この閾値を実験により調整する。

2.7 形態素解析

本研究の最終的な目的は、言語に依存しないツールの構築であるが、実験段階では便宜的に日本語を対象とする。日本語は、英語等の言語と違い、単語間の区切りの位置が明確でない。そのため、どの文字からどの文字が一つの単語であるのかを分析する必要がある。今回は、MeCab[6]という形態素解析ツールを使う。MeCab は、Conditional Random Fields を用いて日本語の単語区切りと、それぞれの単語の品詞を推定するツールである。例えば、「東京都に住む」という文章は、「東京　都　に　住む」という区切り方に分割され、「東京」は名詞、「都」は接尾辞、「に」は格助詞、「住む」は動詞であると、品詞が解析される。

一般に、格助詞や助動詞は機能語であるので、文章の特性を調べる際には重要でない。その一方、名詞、動詞、形容詞は内容語になりやすい傾向にある。本研究では、機能語である格助詞、助動詞等の機能語の除去を MeCab での形態素解析の段階で行う。対象言語を広げる場合には、機能語の除去のための代替の手段を考える必要がある。

3 実験

3.1 実験概要

今回は 2 つの実験を行った。まず一つ目の実験では、tf-idf によりインターネット上のニュース記事と twitter の書き込みから内容語を抽出するシステムを構築し、その性能を評価した。内容語の抽出に関しては、サンプル記事と、その記事に言及している tweet に現れている各単語について tf-idf を計算し、ソートすることにより、内容語と判定する基準である閾値がどのような値になるのかを検討した。次に、ニュース記事と tweet の間の関連度を「単語の一致」のみを指標として計算し、ニュース記事と tweet のアラインメントを取った。すなわち、式 16 の β と γ をゼロにしたモデルでアライントを取ったことになる。結果を人間の手によりアラインメントを取ったものと比較し、アラインメントを取る際の関連度 $\mathrm{Rel}(s,t)$ の閾値を検討する。アラインメントの結果の評価は、precision-recall**??**を尺度に行う。precision と recall は検索エンジンの性能評価などに用いられる評価尺度で、以下の式によって表される。

$$\mathrm{precision} = \frac{R}{N} \tag{17}$$

$$\mathrm{recall} = \frac{R}{C} \tag{18}$$

ここで、N はアラインメントした結果 web ニュースの断片に対応づけられた tweet の数、C は正解データで web ニュースの断片に対応づけられるべきとされた tweet の数、R はアラインメントした結果 web ニュースの断片と対応づけられ、なおかつ正解とも一致した tweet の数である。これらの値を合わせた F-score で、アラインメントの評価を行う。F-score は precision と recall の調和平均であり、以下の式によってあらわされる。

$$F = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} \tag{19}$$

$$= \frac{2R}{N + C} \tag{20}$$

3.2 実験条件

今回の実験で使用するニュース記事は、Twitter 上で”@Google_news_jp”[9]というアカウントによって取り上げられているニュースから無作為に抽出したものを使用する。”@Google_news_jp”は、様々な報道機関が提供しているインターネット上の記事のタイトル・URL を紹介しているアカウントである。今回の実験で使用する tweet は、収集したニュース記事のタイトル、URL を Twitter の公式検索ツールで検索した結果ヒットした tweet をその記事に言及する tweet とみなし、使用する。ただし、”@Google_news_jp”のような単なるニュース紹介のアカウントの tweet は、個人の意見や感想を含まないために除外する。記事に言及する tweet は記事のタイトルや URL を内部に含む場合が多いので、その部分は tweet の本文から除外して分析をする。

tf-idf において idf を計算するためのデータセットとしては、”@Google_news_jp”のニュース記事から 20 記事を集めた物を使用する。また、”@Google_news_jp”のニュース記事から多くの tweet に言及されている記事を二つ[10][11]を選び、内容語の抽出の評価や、アラインメントの評価に使用する。

3.3 実験結果

3.3.1 内容語の抽出

いくつかのニュース記事に対して、その中に出現する全ての単語の tf-idf を計算し、その値を降順にソートして並べた結果を図??に示す。

図??のように、tf-idf の分布が三つの集団に分かれる結果となった。一つは、一番左側の tf-idf が極端に高い単語群である。これは、いわゆるキーワードであると言える。実際、コーヒーの健康への影響を題材にした記事では、「コーヒー」「杯」といった単語がこの集団に属していた。二つ目は、中央部の平らな、平均的な tf-idf をもつ集団である。コーヒーを題材にした記事では、「カフェイン」や「サウスカロライナ」などの内容語が多く所属していたのに対し、「ただし」「よう」等のあまり情報を含まない単語も所属していることが確認された。三つ目は、右側で急降下している、低い tf-idf をもつ集団である。この集団には「て」「に」「を」「は」などの助詞を含む機能語が多く所属していた。

3.3.2 アラインメント

「コーヒーの健康に対する影響」のニュースに対して手動で正解を作り、アラインメントの結果から precision, recall, f-score を算出した結果を図 4 に示す。

[9]https://twitter.com/Google_News_jp
[10]http://www.asahi.com/national/update/0825/TKY201308250154.html
[11]http://zasshi.news.yahoo.co.jp/article?a=20130830-00000003-sasahi-soci

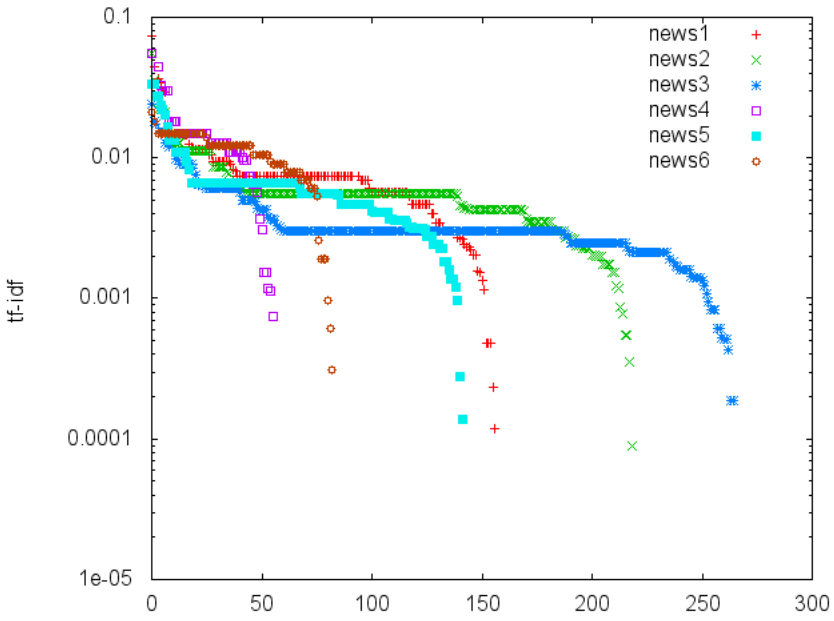

図 3: tf-idf の計算結果

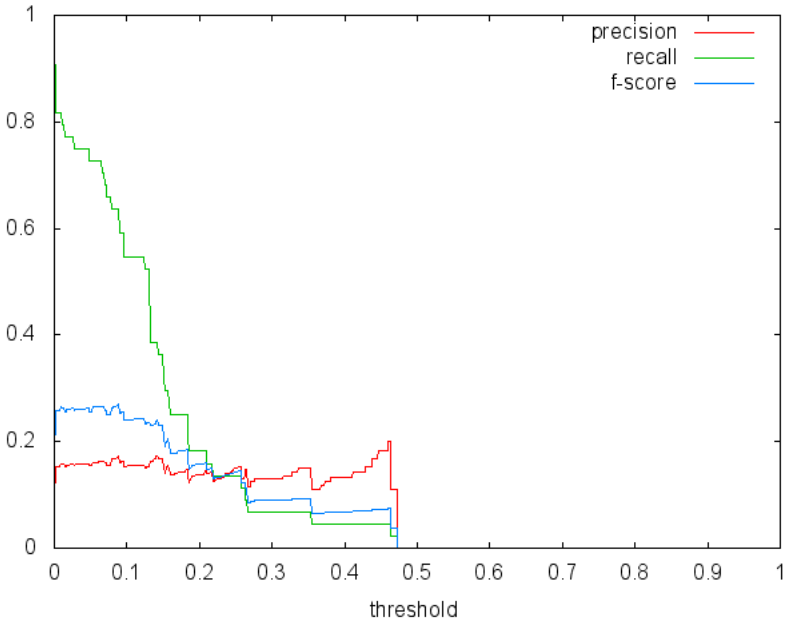

図 4: アラインメントの結果

F 値の最大は、閾値が 0.088 で 0.27 となった。本来、閾値が高くなると recall は低くなり、precision が高くなるはずである。しかし、閾値が高くなっても precision は低いままであった。また、記事中のほとんど全てのセンテンスと対応づけられてしまうような tweet が多く見られた。

4　考察

4.1　内容語の抽出

内容語の抽出では、図??のように値がおおまかに 3 つの部分に分かれる結果となった。このうち一番 tf-idf の低い集団が、格助詞等の意味を持たない単語の集団であることが確認された。この部分に含まれる単語を除外することで内容語の抽出を行うことができる。しかし、この集団と内容語を含む中央部の集団との境界となる tf-idf の値は一定ではなく、分析する文章により異なる。そこで、それぞれの記事の分布からこの境界の値を算出するアルゴリズムを考える必要がある。また、中央部の集団にも意味を持たない機能語が混ざっているので、それをどのように処理するかを考える必要もある。一つの対策として、現在 idf の算出に使用している記事の数が 20 と少ないので、収集するデータ数を増やして idf の精度を上げることが考えられる。

4.2　アラインメント

アラインメントでは、本来、閾値が高くなると recall は低くなり、precision が高くなるはずであるところ、閾値が高くなっても precision は低いままであった。これは、比較的厳選した tweet と news の対応でも適合度が低いということで、まだアラインメントの精度が低いことを表す。これはこの先、類似語や関連語の検出を組み込むことで改善されることが期待される。また、記事中のほとんど全てのセンテンスと対応づけられてしまうような tweet が多く見られたが、そのような tweet の対応づけはユーザーからみて不適切に見える可能性が高いので、関連語・類似語の検出を組み込んだ後にもそのような問題が起こっていないか注意する必要がある。。

5　今後の課題

5.1　データ収集の自動化

今後の課題として、内容語の抽出の精度をあげるために、idf の精度を上げることが挙げられる。そのためには idf 算出のためのサンプルデータを増やす必要がある。現在サンプルデータは手動で収集しているが、これでは限界がある上、ニュースの本文では時代が変わると語彙が代わり、idf がうまく機能しない可能性もある。そこで、idf 算出のためのサンプルデータ収集を自動化する必要がある。手法としては、現在 Twitter 社が tweet の検索 API を提供しているので、”@Google_News_JP”のようなニュース配信アカウントからインターネット上のニュース記事の URL を自動取得し、その URL のウェブページからテキストデータを収集する、というものが考えられる。

同様に、アノテーションをする際にそのニュース記事に言及している tweet を取得することも自動化する必要がある。

5.2　内容語の抽出

今回得られた実験結果をもとに、tf-idf 分布における下位集団の自動除外を行い、内容語の抽出を行う。また、データ集種の自動化により idf の精度向上が実現した際には、rf-idf 分布における平均的集団から機能語を除去する手段についても検討する。

5.3　言い換え語・類似語の検出

今回の実験では未実現となった「言い換えの関係にある語」「互いに類似する語」の検出を実装する必要がある。言い換えの関係にある語は WordNet の Synset を利用して検出する。これは容易に実現可能であると考えられる。問題は互いに類似する語の検出であるが、Dekang Lin の手法 [3]、Latent Space でのモデル化 [4] を実装し、結果を比較検討しながら二つの手法から選択及び組み合わせをして実現する。

5.4　ユーザー評価の実施

上記の課題を解決した後に、実際に Web アプリケーションとして本ツールを実装し、ユーザー評価を行う。ユーザー評価では、10 ほどの記事にアノテーションを施したものをユーザーに評価してもらい、アノテーションは正確であったか、不要なアノテーションはあったか、このツールによりニュース記事は読みやすくなったかなどをアンケートにより調査する。

6　結論

本研究では、ソーシャルネットワーキングサービスである Twitter 上の投稿をインターネット上のニュース記事に自動付与するソーシャルアノテーションツールの構築を行う。要素技術として内容語の抽出、言い換え語の検出、類似語の検出と、それらを利用した短文同士のアラインメントが挙げられる。現段階ではこのうち、内容語の抽出と、言い換え語・類似語を利用しないアラインメントの実装を行った。内容語の抽出に関しては今回得られた結果から良いアルゴリズムを作成する目処を立てる事ができた。アラインメントに関してはまだ精度向上の必要性があるが、実験を通して精度向上のための方針を得ることができた。また、今後の研究で内容語の抽出と言い換え語・類似語の検出を実現することができれば、より高いアラインメント精度の実現が見込めると考える。

ソーシャルアノテーションツールは、情報のプロの書き手と一般人の読み手の距離を縮め、また、難解な情報へのアクセスをより容易にする画期的なツールである。本研究を通して、人の情報との関わり方をより豊かにできるものと自負している。そのために、今後も研究活動により一層尽力して行きたい所存である。

参考文献

[1] 高村大地, 奥村学. 言語処理のための機械学習入門 (コロナ社), 2010

[2] Gerard Salton. Introduction to Modern Information Retrieval (Mcgraw Hill, Inc.), 1986

[3] Dekang Lin. Automatic Retrieval and Clustering of Similar Words. COLING '98 Proceedings of the 17th international conference on Computational linguistics - Volume 2 Pages 768-774

[4] Weiwei Guo, Mona Diab. Modeling Sentences in the Latent Space. ACL '12 Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1 Pages 864-872

[5] Inderjit S Dhilon, Subramanyam Mallela, Dharmendra S. Modha. Information-Theoretic Co-clustering. KDD '03 Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining Pages 89-98

[6] 工藤 拓, 山本 薫, 松本 裕治. Applying Conditional Random Fields to Japanese Morphologiaical Analysis, 情報処理学会研究報告. 自然言語処理研究会報告 2004, Pages 89-96

[7] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schu"tze: Introduction to Information Retrieval, Cambridge University Press, 2008.

[8] Ted Pedersen, Siddharth Patwardhan and Jason Michelizzi. WordNet::Similarity - Measuring the Relatedness of Concepts, HLT-NAACL–Demonstrations '04 Demonstration Papers at HLT-NAACL 2004 Pages 38-41