オンライン百科辞典を対象とする有効期限切れ情報データベースの作成

土屋雅稔　　横井康孝
豊橋技術科学大学大学院 情報・知能工学専攻
{tsuchiya,yokoi}@is.cs.tut.ac.jp

1 はじめに

ウェブを通じて提供される膨大な情報は，我々の日常生活において欠かせないツールとなっている．しかし，全ての情報を参照することは非現実的であり，適切な取捨選択が必要である [1, 2]．特に，ウェブ上の情報は全てが適時に更新されているわけではないため，現在参照するには不適当な情報についての取捨選択は重要である．具体例として，図 1 のような情報を考える．文 s_b に記載されている通り大志田駅は 2016 年に廃止されているため，文 s_t に基づいて盛岡市浅岸字大志田に行く交通手段を考えることは適切ではない．

ウェブには，文 s_t のような情報が多数存在しており，情報利活用の障害となっている．このような情報の利活用を支援するため，情報を時系列に注目して整理・提示する方法が提案されている [3, 4]．しかし，先行研究では，利用者の観点から現在参照するには不適当な情報の自動判定は行っていない．

本稿では，文 s_t のような情報を**有効期限切れ情報**と呼び，有効期限切れ情報を自動判定するタスクを提案する．有効期限切れ情報の自動判定タスクを設定するにあたっては，判定対象となる文 s_t および根拠となる文 s_b の 2 文を入力して文 s_t が有効期限切れであるか否かを判定するタスク設定と，判定対象となる文 s_t のみを入力して文 s_t が有効期限切れであるか否かを判定するタスク設定の 2 通りが考えられる．有効期限切れ情報の自動判定によって情報利活用を支援するという，有効期限切れ情報の自動判定タスクが目指す本来の目的から考えると，文 s_t のみを入力とする後者のタスク設定が自然である．しかし，後者のタスク設定を実現するためには，文 s_t が有効期限切れである根拠を発見するというタスクを同時に解く必要があり，タスクが極端に難しくなることが予想される．そこで本稿では，判定対象となる文 s_t および根拠となる文 s_b の 2 文を入力として，文 s_t が有効期限切れであるか否かを判定するという前者のタスク設定に限定して検討する．以後，判定対象となる文 s_t を**判定対象文**，根拠となる文 s_b を**基準文**と呼ぶ．

s_t	大志田駅は，2010 年現在，岩手県盛岡市浅岸字大志田にある無人駅であり，1 日 3 本の電車が停止する．
s_b	大志田駅は，2016 年に廃駅となった岩手県盛岡市浅岸字大志田の無人駅で，2018 年現在，使用されていない．

図 1　有効期限切れ情報の例

有効期限切れ情報を収集するテキストとしては，オンライン百科事典 (日本語版 Wikipedia) を利用する．オンライン百科事典は継続的に更新されており，旧版のオンライン百科事典と，新版のオンライン百科事典を比較すると，追加された記事や，記述内容が更新された記事が見つかる．特に，ある項目に関する記事が更新されている場合，旧版の記述には有効期限切れ情報が含まれ，かつ，新版の記述には判定根拠となる情報が含まれることが多いと期待される．

以下，有効期限切れ情報の分析と定義および有効期限切れ情報データベースの作成方法について述べる．

2 有効期限切れ情報の定義

情報の有効期限切れとは，直感的には，図 1 のように，古い説明文 s_t の情報が新しい説明文 s_b によって上書きされ，古い説明文の情報が有効性を失っている状態である．ここで，以下のような文 s_n を考える．

> s_n　盛岡市が大志田駅から浅岸駅までの 13 キロの近郊自然歩道を整備し、2015 年 5 月 15 日から利用可能となった。

3 つの文 s_t, s_b, s_n は全て，大志田駅について述べた

文である．しかし，文 s_b を基準として文 s_t が有効期限切れと言うことは可能と考えられるのに対して，文 s_n を基準として文 s_t が有効期限切れと言うことは困難である．この例は，有効期限切れ関係は，あらゆる文対に対して成り立つわけではなく，ある特定の文対に対してのみ成り立つことを示唆している．

そこで，この節では，有効期限切れについて，含意関係との関係に注目しつつ，形式的な定義を与えることを試みる．

2.1 有効期限切れ現象の検討

最初に，有効期限切れが実際に起きている例文を参照しながら，有効期限切れとは，どのような現象であるのかを検討する．

有効期限切れが起きている例として，以下のような 2 つの文 s_1, s_2 を考える．

s_1 = 日本で最も高いビルは，2010 年現在，横浜ランドマークタワーである

s_2 = 日本で最も高いビルは，2015 年現在，あべのハルカスである

「日本で最も高いビル」に関する文 s_1 の情報は，文 s_2 によって上書きされており，2020 年現在の利用者視点では，文 s_1 の情報は有効期限切れである．

形式的に議論するため，文 s の記述内容を，文 s が注目している時間 t と，時間とは関係なく成立する記述内容 x の 2 つ組として考える．

$$s = \langle x, t \rangle$$

まず，文 s_1 の記述内容を，2 つ組 $\langle x_1, t_1 \rangle$ に分解する．

$s_1 = \langle x_1, t_1 \rangle$

x_1 = 日本で最も高いビルは，横浜ランドマークタワーである

t_1 = 2010 年

文 s_2 の記述内容も，同様に 2 つ組 $\langle x_2, t_2 \rangle$ に分解する．

$s_2 = \langle x_2, t_2 \rangle$

x_2 = 日本で最も高いビルは，あべのハルカスである

t_2 = 2015 年

ここで，文 s_1 の時間情報のみを t_2 に置き換えた文 $s_{1,2}$ を考える．

$s_{1,2} = \langle x_1, t_2 \rangle$

= 日本で最も高いビルは，2015 年現在，横浜ランドマークタワーである

この時，文 $s_{1,2}$ と文 s_2 は，矛盾している．

次に，有効期限切れを引き起こさない例として，以下のような文 s_3 を考える．

s_3 = 世界で最も高いビルは，2015 年現在，ブルジュ・ハリファである

文 s_1 は日本で最も高いビルについて述べている．それに対して，文 s_3 は世界で最も高いビルについて述べており，日本で最も高いビルについてはまったく言及していない．そのため，文 s_3 が基準となって，文 s_1 が有効期限切れとなるとは考えられない．また，文 $s_{1,2}$ と文 s_3 は，矛盾していない．

以上の観察より，本稿では，有効期限切れを，矛盾を手がかりとして定義する．

2.2 有効期限判定と含意関係判定の関係

本稿では，2.1 節で述べた通り，情報の有効期限切れを，矛盾を手がかりとして定義する．以下では，その形式的な定義について述べる．

最初に，含意関係判定を定義する．先行研究 [5] にしたがうと，含意関係判定は，2 つの文 s_h, s_p が与えられた時，その 2 つの文の含意関係に応じて以下の 3 通りの値を返す関数 f と見なせる．なお，本稿では，含意関係の方向については考慮せず，どちらかの方向の含意関係が成り立っている時は含意であるとする．

$$f(s_h, s_p) = \begin{cases} \text{if } s_h \rightarrow s_p \vee s_p \rightarrow s_h \\ \quad \text{含意} \\ \text{if } s_h \wedge s_p = \phi \\ \quad \text{矛盾} \\ \text{otherwise} \\ \quad \text{中立} \end{cases} \tag{1}$$

次に，有効期限判定を定義する．まず，判定対象文 s_t と基準文 s_b を，それぞれ内容と時間情報の 2 つ組に分解する．

$$s_t = \langle x_t, t_t \rangle \tag{2}$$

$$s_b = \langle x_b, t_b \rangle \tag{3}$$

判定対象文 s_t と基準文 s_b が与えられた時，2.1 節の観察に基づいて，有効期限判定を，以下の 2 通りの値を返す関数 g として定義する．

$$g(s_t, s_b) = \begin{cases} \text{if } f(s_t, s_b) = \text{矛盾} \\ \quad \vee f(\langle x_t, t_b \rangle, s_b) = \text{矛盾} \\ \quad \vee f(s_t, \langle x_b, t_t \rangle) = \text{矛盾} \\ \quad \text{期限切れ} \\ \text{otherwise} \\ \quad \text{期限内} \end{cases} \tag{4}$$

以上のように，本稿では，情報の有効期限切れを，基準となる情報が存在して初めて決定できる概念として扱う．

情報の有効期限切れを，以上のように定義すると，多くの場合に利用者の直感とよく一致するが，時間変化する情報の有効期限切れについては注意が必要である．例として，以下の 2 つの文 s_4, s_5 を考える．

s_4 = 東都大学の 2014 年度の受験者数は 2000 人である

s_5 = 東都大学の 2015 年度の受験者数は 3000 人である

文 s_4 を，以下の 2 つ組に分解する．

$s_4 = \langle x_4, t_4 \rangle$

x_4 = 東都大学の受験者数は 2000 人である

t_4 = 2014 年度

文 s_5 も，同様に 2 つ組に分解する．

$s_5 = \langle x_5, t_5 \rangle$

x_5 = 東都大学の受験者数は 3000 人である

t_5 = 2015 年度

ここで，文 s_4 の時間情報のみを t_5 に置き換えた文 $s_{4,5}$ を考える．

$s_{4,5} = \langle x_4, t_5 \rangle$

= 東都大学の 2015 年度の受験者数は 2000 人である

この時，文 $s_{4,5}$ と文 s_5 は矛盾している．そのため，本稿では，文 s_4 は 2014 年度の情報としては未だに真であるにも関わらず，式 4 の定義に基づいて，文 s_4 は文 s_5 によって上書きされており，文 s_4 の情報は有効期限切れであると見なす．言い換えれば，本稿の定義では，有効期限を現時点 (正確には，基準テキストが記述された時点) の情報として有効であるか否かを考えていることになる．

さらに，文 s_5 の代わりに，以下のような文 s_6 を考える．

s_6 = 東都大学の 2015 年度の受験者数は，昨年度に比べて 1000 人増加した

文 s_6 は，以下のように 2 つ組に分解できる．

$s_6 = \langle x_6, t_6 \rangle$

x_6 = 東都大学の受験者数は，昨年度に比べて 1000 人増加した

t_6 = 2015 年度

文 s_4 と文 s_6 の対，または文 s_5 は，同じ事実「東都大学の 2015 年度の受験者数は 3000 人である」についての 2 通りの異なる表現である．にも関わらず，文 $s_{4,6}$ は文 s_6 とは矛盾しないため，式 4 の定義に従うと，文 s_6 を基準として考える場合には，文 s_4 は有効期限切れではない．言い換えれば，有効期限切れの判定にあたっては，2 つの文から演繹される結果までは考慮することなく，その 2 つの文に基づいて考えるという立場をとる．

3 有効期限切れ情報データベースの作成

3.1 対象テキストの選定

有効期限切れ情報の自動判定によって情報利活用を支援するという，有効期限切れ情報の自動検出タスクが目指す本来の目的から考えると，判定対象文としては，個人によるブログ記事に含まれる文のような自由な文体の文を，基準文としては，新聞記事などの信頼性が高く定型的な文体の文を選択することが適切と考えられる．この場合，判定対象文はブログ記事から，基準文は新聞記事から収集することになる．しかし，膨大なブログ記事から有効期限切れ情報を含む判定対象文を収集することが困難であるだけでなく，その判定対象文に対応する根拠となる基準文を収集することも困難である．

このような文対を大量に作成する方法としては，判定対象文 (または基準文) を多数の作業者に提示し，提示した文に適するような基準文 (または判定対象文) を作成するように依頼するクラウドソーシングが広く用いられている [6]．しかし，この方法で作成されたデータには，作業者によって不適切なバイアスが混入することが指摘されている [7, 8]．

そのため，本稿では，オンライン百科事典の継続的な更新という性質に注目し，更新時期が古い記事から判定対象文を，更新時期が新しい記事から基準文を収集する．

3.2 アノテーション対象差分の抽出

本稿では，有効期限切れ情報データベースの作成対象として，オンライン百科事典である日本語版 Wikipedia に注目する．2014 年 12 月 11 日時点のダンプデータと 2018 年 12 月 20 日時点のダンプデータを対象として，WikiExtractor[1)] を用いたプレーンテキスト化を行い，記事数の変化を調査した．調査結果を，表 1 に示す．本稿では，双方に存在する記事 936,985 件を対象とする．

双方の日時に存在する記事について，記事の記述内容を比較すると，さまざまな差分が発見される．まず，本稿では，問題を単純化するために，編集によって変化している文が 1 文であるような差分を対

1) https://github.com/attardi/wikiextractor

表 1　日本語版 Wikipedia の記事数の変化

2014 年 12 月 11 日時点の記事数	943,488
2018 年 12 月 20 日時点の記事数	1,132,813
追加記事数	195,828
削除記事数	6,503
双方に存在する記事数	936,985

象とする．また，有効期限判定の対象として適切な差分を収集するには，時間経過によって変化した情報に言及している差分を収集する必要がある．そのため，2014 年 (判定対象文 s_t の記述時点) ら 2018 年 (基準文 s_b の記述時点) までの日時に対する言及を含む差分や，同じ期間に追加された記事に対する言及を含む差分を対象とする．このように考えると，193,142 箇所の差分が候補となる．

193,142 箇所の差分には，句点の追加 · 削除，誤字脱字の修正など，微細な編集のみが行われている場合も非常に多く含まれている．本稿の目的に照らすと，判定対象文と基準文で記述内容が本質的に変化している場合を対象として，収集およびアノテーションを行いたい．そこで，判定対象文 s_t と基準文 s_b の最長共通部分列 $LCS(s_t, s_b)$ が，文 s_t, s_b に占める割合が 0.6 以下であるような差分のみを対象とする．このように考えると，15,648 箇所の差分が候補となる．

最後に，できるだけ良質な記事の差分を優先して収集およびアノテーション対象とすることを考える．そのため，SQuAD[9] と同様に，Wikipedia の記事間のリンクに基づく PageRank[10] の上位記事から収集した 9,600 箇所の差分を対象とする．

3.3　アノテーション

3.2 節で説明した手順で抽出した 9,600 箇所の差分に対して，3 種類の含意関係ラベル (含意，矛盾，中立) と 2 種類の有効期限ラベル (期限切れ，期限内) を人手で付与した．付与したラベルの内訳を，表 2 に示す．表 2 より，32.2%(3,090 箇所) の差分が有効期限切れと判定された．この判定結果から，日本語版 Wikipedia には有効期限切れ情報が多く存在することがわかり，有効期限切れ情報の自動判定が必要であることが示された．

また，表 2 より，含意関係ラベルが含意または中立である場合であっても，有効期限切れと判定される差分も多く存在していること，すなわち，従来の含意関係判定と本稿が提案する情報の有効期限判定には違いがあることがわかる．以下に，含意関係ラベルが含意，かつ，有効期限ラベルが期限切れとなる差分の実例を示す．

表 2　含意関係ラベルおよび有効期限ラベルの分布

	期限切れ	期限内	合計
含意	430　(4.5%)	3,765 (39.2%)	4,195　(43.7%)
矛盾	1,359 (14.2%)	0　(0.0%)	1,359　(14.2%)
中立	1,301 (13.6%)	2,745 (28.6%)	4,046　(42.1%)
合計	3,090 (32.2%)	6,510 (67.8%)	9,600 (100.0%)

s_7 = 改札は、2013 年 1 月 13 日をもって高架下の 1 ヶ所に統合された。

s_8 = 改札は、2013 年 1 月 13 日をもって高架下の 1 ヶ所に統合されたが、2016 年 4 月 24 日に西側に nonowa 口（IC カード専用）が新設され、2 ヶ所となった。

この 2 文 s_7, s_8 は，国立駅に関する記事の一部である．文 s_7 で説明されている事実は，文 s_8 でも同様に説明されているため，文 s_8 から文 s_7 に対する片方向含意は成り立っていると考えられる．しかし，文 s_7 における「改札は高架下の 1 ヶ所」という説明は，文 s_8 によって「改札は高架下と西側に 2 ヶ所」と上書きされているため，有効期限ラベルは期限切れとなる．

3.4　作業者間一致度

アノテーションが安定して行われているかどうかを検討するため，作業者間一致度を調査する．3.2 節で述べた Wikipedia の記事間リンクに基づく PageRank の上位記事から，300 箇所の差分を調査対象とし，2 名の作業者による独立したアノテーション作業を実施した．2 名の作業者による含意関係判定の一致率は 83.7%(251 箇所)，有効期限判定の一致率は 88.0%(264 箇所) だった．作業者間で判断が異なった箇所について精査したところ，文外知識の理解度などによるものが多く見られた．

4　結論

本稿では，有効期限切れ情報の自動判定タスクを提案した．有効期限切れ現象について分析を行い，分析に基づいて，情報の有効期限切れを含意関係判定と関連付けて定義した．有効期限切れ情報データベースを，オンライン百科事典の継続的な更新という性質に基づいて作成する方法について述べた．

参考文献

[1] Catherine C. Marshall and Frank M. Shipman. Spatial hypertext and the practice of information triage. In *Proceedings of the Eighth ACM Conference on Hypertext*, HYPERTEXT '97, pp. 124–133, New York, NY, USA, 1997. Association for Computing Machinery.

[2] Sofus A. Macskassy and Foster Provost. Intelligent information triage. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pp. 318–326, New York, NY, USA, 2001. Association for Computing Machinery.

[3] Gabriel Pui Cheong Fung, Jeffrey Xu Yu, Huan Liu, and Philip S. Yu. Time-dependent event hierarchy construction. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, p. 300–309, New York, NY, USA, 2007. Association for Computing Machinery.

[4] Canhui Wang, Min Zhang, Liyun Ru, and Shaoping Ma. Automatic online news topic ranking using media focus and user attention based on aging theory. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, p. 1033–1042, New York, NY, USA, 2008. Association for Computing Machinery.

[5] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 216–223, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

[6] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[7] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112, June 2018.

[8] Masatoshi Tsuchiya. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[9] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016.

[10] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the we.b. Technical Report 1999-66, Stanford InfoLab, November 1999.