コーパスへのアノテーションとガイドライン統合的な管理手法

大内田賢太 † 金進東 ‡ 高木利久 †* 辻井潤一 ‡§
† ライフサイエンス統合データベースセンター
‡ 東京大学情報理工学系研究科コンピュータ科学専攻
* 東京大学大学院理学系研究科新領域創成科学研究科 情報生命科学専攻
§School of Computer Science, University of Manchester
oouchida@dbcls.rois.ac.jp, jdkim@is.s.u-tokyo.ac.jp, takagi@dbcls.rois.ac.jp, tsujii@is.s.u-tokyo.ac.jp

1 はじめに

一般的に、人手によるコーパスアノテーションにおいて一番の問題とされるのは、付記 (アノテーション) の一貫性の維持である。付記の一貫性を維持するために、アノテーターは付記作業中、アノテーションガイドラインを作成し共有する必要がある。ガイドラインの一部分は付記のプロジェクトの初期段階で作成されるが、それ以外の部分は付記作業の過程の上で作成される。

ガイドラインには、アノテーターがどのように付記していいのか迷うような、ボーダーライン上のケースを扱う方法について記述されてある。しばしばガイドラインは付記作業の前に完全なガイドラインを用意しておくことは不可能に近い。それは付記される情報が複雑になるほど高くなる。ガイドラインにはどのように付記していいのか迷うような、ボーダーライン上のケースを扱う方法についての記述が含まれている。アノテーターもしくはプロジェクト管理者はそのケースに題しての指針を作成し既存のガイドラインに追加もしくは既存のガイドラインを修正する必要がある。こんなように付記作業とガイドライン作成の作業は共に行われる場合が多い。ガイドラインは、付記作業の一貫性の維持のために必要だけでなく、後に付記されたコーパスを利用する際にもコーパスの理解を深めるために用いられる。しかし、ガイドラインの作成に関する研究はこれまであまりなされてきていない [5]。

本論文では、付記されたコーパスと共にガイドラインを管理する枠組みを提案し、ガイドラインとコーパス間での相互参照をサポートする。また、既存の付記ツールである XConc Suite[1] にプラグイン可能な提案手法をサポートする **GuideLink** の実装を行った。

2 関連研究

2.1 コーパスアノテーションツール

広く知られているアノテーションツールとして WordFreak [2]、MMAX [7]、Knowtator [9]、GATE [1]、XConc Suite などがある。MMAX は、マルチレベルでの付記に適したアノテーションツールである Knowtator は、オントロジーをベースとした付記作業に適したツールである。GATE は付記作業の基盤システムである。WordFreak と XConc Suite は様々な形式のコーパスや付記に対応することに主眼を置いている。また、機械学習の技術を取り入れ、人手による付記作業量を減らす手法も提案されてきている [11]。しかし、筆者の知りうる限り、アノテーションガイドラインの管理作業をサポートするアノテーションツールは存在しない。

2.2 アノテーションガイドラインとその管理作業

現在、ガイドライン管理作業に関する研究があまり行われていないが、ガイドラインの文章化の重要性の認識は広まっている。最も有名なアノテーションコーパスの一つである Penn Treebank [6] でもまた、ガイドラインは文章化されて共有されている。他には、SUSANNE corpus [10] のように書籍化されて共有される例もある。

近年の付記作業では、より一般的な管理手法によってガイドラインが管理されている。例えば、the Caderige project [3] では、メールによりアノテーター間の意思疎通が取られ、そのメールのアーカイブがガイドラインとして保存されている。PennBioIE [4] では、ガイドラインをまとめた特設の Web ページを作成・更新することで意思疎通を取っている。GENIA では、Wiki システムを用い情報共有を行っている。[2]

メールや Web ページを用いた手法は、ガイドラインの文章化および共有・検索において有用だが、ガイドラインの管理作業を行うシステムと付記作業を行うシステムを別々に用意する必要がある。最も大きな問題としてガイドラインは付記の実例とつながる事によってその価値が上がるが、一般的な文書化システムをガイドライン作成に使うことになるとそのつながりが実例の文書への copy-paste で行われるため、作成のコストも高いし後参照も不便になる、との問題が生じる。

3 コーパスアノテーションの流れ

図 1 は、筆者が付記従事者と議論し、コーパスアノテーションの一般的な流れをモデル化したものである。このモデルでは、付記対象となる単語列が与えられたとき、アノテーターがどのように判断して付記を行うかを示している。もしこの判断が容易に行われるものであれば、アノテーターはガイドラインの手助けを必要とせず付記を行うことができる。この判断が容易に行われるものでなければ、アノテーターはガイドラインを参照することになる (2)。もし、このとき参考にな

[1] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=XConc+Suite+User+Manual

[2] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=GENIA+corpus

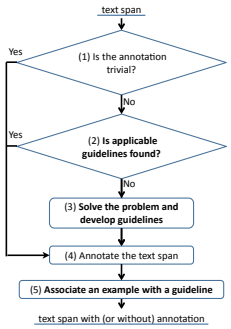

図 1: ガイドラインを考慮した付記の流れ

るガイドラインを見つけることができた場合はガイドラインを参考に付記を行うことができるが、見つけることができなかった場合、アノテーターは自ら判断し付記をどのように行うかを決める必要がある。その判断と判断の理由は、ガイドラインとして整理され保存し (3)、付記作業を行う (4)。付記が行われた単語列は、付記された情報とセットとしてアノテーションインスタンスと呼ばれ、付記作業の情報を多く含んでいる。これらのアノテーションインスタンスは、適切なガイドラインに関連づけられて、ガイドラインの良い具体例として参考にされる (5)。

一般的なアノテーションツールでは、ガイドライン管理を考慮してなく、モデル中の (4) のステップしかサポートされていない。そのため、ガイドラインを管理する作業は必ず、アノテーションツールによる付記作業とは別に行う必要がある。本論文では、本章で説明したコーパスアノテーションのモデルを基に、付記管理作業と付記作業を統合化した付記のフレームワークを提案する。

4 3 レイヤーモデル

ガイドラインの管理作業を付記作業のフレームワークに統合するために、我々はデータ構造として 3 レイヤーモデルを提案する。ガイドラインの管理についてまでサポートしてない一般的な多くのアノテーションツールでは、テキストレイヤーとアノテーションレイヤーで構成される、2 レイヤーモデルしかサポートしていない。ガイドラインの管理のためには、さらにアノテーションガイドラインレイヤーを加えて、3 レイヤーモデルをサポートする必要がある。この章では、3 レイヤーモデルのデータ構造 (図 2) について説明する。

テキストレイヤーでは付記対象となるテキストドキュメントを管理する単語列はそのコーパスの先頭文字からの文字数によって管理される。アノテーションレイヤーは、テキストドキュメント上のどの場所に付記するか管理する。コーパスアノテーションは、テキストレーヤーで管理されているコーパスとアノテーションレイヤー上の情報を関連づける作業と定義することができる。アノテーションインスタンスは一般的に、単語列と記述子の 2 対によって表現される。アノテーションインスタンス中の単語列は、テキストレイヤー中の

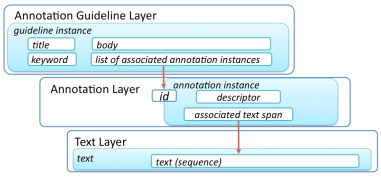

図 2: 3 レイヤーモデルのデータ構造

テキストドキュメントの特定の範囲を指し示す。そのため、アノテーションレイヤーはテキストレイヤーに依存している。また、記述子は、単語列に対して付記された言語情報を示す。この記述子は、一般的に、付記作業の事前に用意される。

ガイドラインレイヤーは、アノテーションガイドラインを管理するレイヤーである。ガイドラインのインスタンスはタイトル、本文、キーワード、関連付けられたアノテーションインスタンスのリストで構成されている。タイトルと本文は、一般的にガイドラインの要約と詳細になる。キーワードは、参考となるガイドラインへの参照をサポートするための索引になる。関連付けられたアノテーションインスタンスのリストは、ガイドラインからコーパス上の実例を参照するのに役立つ。アノテーターは、タイトルや本文、キーワード、関連付けられたアノテーションインスタンスを基に、参考となるガイドラインを探し出すことができる。

さらに、実際の付記作業においては、付記作業でタグが付けられなかった単語列にも、付記の判断における多くの情報が含まれている場合がある。我々は、このような単語列によって構成されるアノテーションインスタンスを、ネガティブアノテーションインスタンスと呼ぶこととする。しかし、一般的な付記のフレームワークでは、記述子が付けられた無い単語列をアノテーションインスタンスとして扱うことができない。

ネガティブアノテーションインスタンスは、一般的なアノテーションツールではサポートしていないが、ガイドラインの具体例として有用となる場合が存在する。ネガティブアノテーションインスタンスをサポートするために、アノテーションレイヤーを拡張する。拡張アノテーションレイヤーでは、アノテーションインスタンスは 3 対の要素 (単語列、記述子、正負判定) によって表現される。実際のコーパス上では、正負判定で正例と判断されたアノテーションインスタンスの単語列には記述子が付けられ、負例と判断された単語列には記述子が付けられない。これにより、ネガティブアノテーションインスタンスを拡張アノテーションレイヤーでサポートすることができ、ガイドラインと関連づけることができるようになった。

5 GuideLink

我々は、前章までに提案したアノテーションフレームワークを基に、ガイドライン管理システムである GuideLink の実装を行った。GuideLink は、4 章で紹介した 2 レイヤーモデルをサポートするアノテーションツール、XConc Suite のプラグインとして実装を行っ

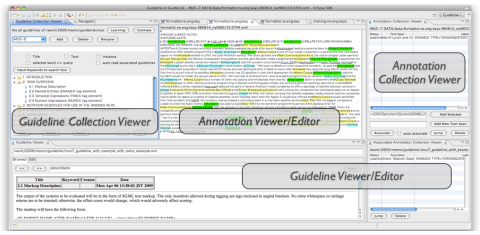

図 3: XConc Suite + GuideLink

た。GuideLink は、ガイドラインレイヤーと拡張されたアノテーションレイヤーをサポートする。GuideLink と XConc Suite を用いることで、アノテーターは二つの異なるシステムを切り替えながら付記作業とガイドライン管理作業を行う必要が無くなった。GuideLink は、図 1 の (2)、(3)、(5) をサポートする。

図 3 は、GuideLink と XConc Suite のスナップショットである。XConc Suite に組み込まれた GuideLink は、次の 4 つのコンポーネントで構成されている。Annotation Collection Viewer は、アノテーションインスタンスの一覧を表示する。Annotation VIewer/Editor は、テキストドキュメント上の単語列を読んだり、付記を行うことができる。Guideline Collection Viewer は、アノテーションガイドラインの一覧を表示する。Guideline Viewer/Editor は、XConc Suite の動作により、アノテーションガイドラインを読んだり、編集したいすることができる。提案手法の目標としては、付記されたコーパスとガイドラインが互いにアクセスしやすい環境を整えておくことになる。

アノテーションガイドラインからアノテーションインスタンスへアクセスする方法は、Guideline Viewer/Editor と Annotation Collection Viewer でサポートされている。Guideline Viewer/Editor によってガイドラインを参照しているとき、Annotation Collection Viewer には、ガイドラインに関連付けられたアノテーションインスタンスの一覧が表示されている。このようにして、アノテーターはアノテーションガイドラインから、ガイドラインにとって具体例と言えるアノテーションインスタンスの参照ができる。

アノテーションインスタンスからアノテーションガイドラインを参照する方法は、Guideline Viewer/Editor や Annotation Collection Viewer、the Guideline Collection Viewer でサポートされている。アノテーターがアノテーションインスタンスからガイドラインを参照したいとき、あるいはアノテーションインスタンスに関連付けられているガイドラインを見たいとき、アノテーターは Annotation Collection Viewer 上のアノテーションインスタンスを選択すれば、Guideline Viewer/Editor にガイドラインが表示される。一方で、この手法では、まだタグが付けられてない、もしくはガイドラインと関連付けられてない単語列から、参考になるガイドラインへの参照を行うことができない。

我々は、ガイドラインに関連付けられたアノテーションインスタンスと、付記対象としている単語列の類似度から、参考となるガイドラインを推測する手法を実装した。類似度を計算する方法として、我々は Weka [12] に実装された support vector machine (SVM) や k-nearest neighbor (KNN) を用い、分類機の学習を行えるようにした。学習データは、ターゲットとする単語と前後 3 つの単語に注目し、素性ベクトルを取ることにした。各単語から品詞情報を獲得するために、OpenNLP tools[3]を使用した。

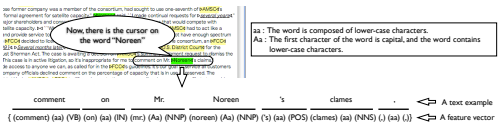

図 4: コーパスからの特徴ベクトル抽出

図 4 はコーパス上のテキスト *"Noreen"* から素性ベクトルを取ってきた例である。例では、*"Noreen"* とその周辺の単語 *"to comment on Mr. Noreen's claims,"* から素性ベクトルを作成している。このコーパスから得られた素性ベクトルを基に、類似度の高い具体例を持つガイドラインを自動的に提案する。

本手法の目的は、付記作業とガイドライン管理を統合したフレームワークの提案を行い、付記されたコーパスとガイドラインとの間のアクセシビリティを向上させることである。本章では、2 種の評価を行う。最初の評価では、実際の付記の流れの中において、本手法の利点がどこにあるか明確にする。既存の手法と我々のフレームワークを用いた手法との比較を行い、その違いを明確にする。次の評価では、適切なガイドラインへの参照方法についての評価を行う。我々は、特定の素性情報を用い、類似度を用いたガイドライン提案手法の評価を行う。

5.1 一般的手法と提案手法との比較

表 1 では、一般的な付記と提案手法による付記の比較を行っている。各ステップの番号は、フローチャートの図 1 と関連付けられている。ステップ (2)、(3)、(5) は主に、ガイドラインを管理・参照するステップである。

一般的な手法では、ガイドラインを管理する際に特定のシステムを用いることは少なく、ワードプロセッサーや Wiki などの一般的なシステムをもちいることが多く、それらのシステムに依存した検索手法、編集手法によってガイドラインを参照、管理することになる。提案手法では、ガイドラインを付記作業と統合した形で管理することができる。

5.2 類似度によるガイドライン提案手法の評価

本手法を検証するため、我々は、MUC-7 named entity annotation [8] のガイドラインを、GuideLink の 7 6 個のガイドラインとして人手により複製した。また、MUC-7 のガイドラインには直接、付記の具体例が書かれてあるため、その具体例を分類機を学習する際の正例として用いることとする。また学習の負例としては、他のガイドラインに記述されている具体例と、MUC コーパ

[3]http://opennlp.sourceforge.net/

表 1: 既存手法と提案手法の比較

図 1	一般的な手法	提案手法
(2)	ワードプロセッサーや Wiki の検索機能を用い、ガイドラインを参照する。	キーワードを用いた検索や、類似度によるガイドライン提案手法を用いた検索を行う。
(3)	例えば、ワードプロセッサーや Wiki などを用い、ガイドラインを更新する。ガイドラインは一般的に、平文で記述される。	ガイドラインレイヤー上にガイドラインにを追加する。そのガイドラインは構造的なデータによって記述される。
(5)	関連するガイドラインの中に、具体例を書き入れる。	ガイドラインからアノテーションインスタンスにリンクを張る。

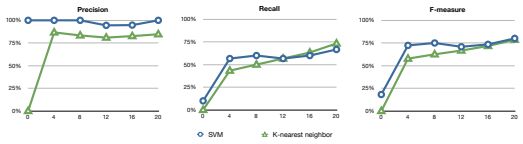

図 5: 自動ガイドライン提案手法のシミュレーション

スからランダムに 1000 個の単語列を取ってきて、それをもちいることとした。本手法における予測される問題として、正例が少なすぎるということが挙げられる。

そこで我々は、各ガイドラインに対して、人手により正例を追加してみる試みを行ってみた。実験では、3 つのガイドラインを選択し、分類機によって正しいガイドラインが提案出来るかどうか評価を行った。(A.3.1[4]、B.3[5]、C.1[6])。評価として、40 個の単語列 (うち 10 個は、分類機によってそれぞれ A.3.1、B.3、C.1 が提案されて欲しい単語列。のこりの 10 個はどのガイドラインも提案されて欲しくない単語列。) に対して分類機の評価を行った。

図 5 は、ガイドライン提案手法において、正例の数を徐々に増やしたときの、Precision、Recall、F-measure の推移を表している。横軸が追加した正例の数、縦軸が各スコアになっている。

図 5 から、正例を増やしていくにつれて、ガイドライン提案手法の精度が上がっているのがわかる。KNN による提案手法では Recall が良いのに対し、SVM による提案手法では Precision が良いことが分かる。今回の実験は限定された条件下、少数のガイドラインによってで行われたが、最終的な結果としては、F-masure が 80% に到達する等、高い可能性を持っていることを示すことができた。

6 まとめ

アノテーションガイドラインの管理は、の一貫性を保つことや、得られたアノテーションコーパスを解析するために有用であることは知られていたが、ガイドラインの管理作業が作業とは別に行われているため、二つの作業を別々に行う必要があった。本論文では、ガイドラインの管理作業を作業と統合化するためのフレームワークを提案した。評価として、本手法を実装したアノテーションシステムにおいて参考となるガイドラインが容易に参照出来ることの検証を行った。

参考文献

[1] GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications (2002)

[2] WordFreak: An Open Tool for Linguistic Annotation (2003)

[3] Alphonse, E., Aubin, S., Bisson, G., Hamon, T., Lagarrigue, R., Nazarenko, A., pierre Manine, A., N??dellec, C., Ould, M., Vetah, A., Poibeau, T., Weissenbacher, D.: Event-based information extraction for the biomedical domain: The caderige project. In: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications. (2004 (2004)

[4] Kulick, S., Kulick, S., Bies, A., Liberman, M., M, M., Winters, S., White, P.: Integrated annotation for biomedical information extraction (2004)

[5] Lu, Z., Bada, M., Ogren, P., Cohen, K., Hunter, L.: Improving biomedical corpus annotation guidelines. In: Proceedings of the Joint BioLINK and 9th Bio-Ontologies Meeting, pp. 89–92 (2006)

[6] Marcus, M., Santorini, B., Marcinkiewicz, M.: Building a large annotated corpus of english: The penn tree bank. Computational Linguistics **19**(2), 313–330 (1993)

[7] Mueller, C., Strube, M.: Mmax: A tool for the annotation of multi-modal corpora. In: In Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems, pp. 45–50 (2001)

[8] N. Chinchor, P.R. (ed.): MUC-7 Named Entity Task Definition (version 3.5) (1997). URL http://www.itl.nist.gov/iad/894.02/related_projects/muc/proceedings/ne_task.html

[9] Ogren, P.V. (ed.): Knowtator: a plug-in for creating training and evaluation data sets for biomedical natural language systems (2006)

[10] Sampson, G.: English for the Computer: The SUSANNE Corpus and Analytic Scheme. Computational Linguistics **28**(1), 102–103 (2002)

[11] Tsuruoka, Y., Tsujii, J., Ananiadou, S.: Accelerating the annotation of sparse named entities by dynamic sentence selection. BMC Bioinformatics **9**(Suppl 11), S8 (2008)

[12] Witten, I.H., Frank, E. (eds.): Data Mining: Practical machine learning tools and techniques (Second Edition). Morgan Kaufmann (2005)

[4] A.3.1 Titles vs. Generational Designators (Titles such as “Mr.” and role names such as “President” are *not* considered part of a person name. However, appositives such as “Jr.,” “Sr.,” and “III” *are* considered part of a person name.)

[5] B.3 Temporal Expressions Containing Adjacent Absolute and Relative Strings (When a time expression contains both relative and absolute elements, the entire expression is to be tagged. The following examples illustrate some of the ways in which elements of relative and absolute time expressions may combine to form taggable time expressions.)

[6] C.1 Scope of Numeric Expressions (The entire string expressing the monetary or percentage value is to be tagged.)