アノテーションガイドラインの管理を行う半自動的アノテーションシステムの提案

大内田賢太† 金進東† 辻井潤一†‡§
†東京大学 情報理工学系研究科
‡School of Computer Science, University of Manchester
§National Centre for Text Mining, University of Manchester
{oouchida, jdkim, tsujii}@is.s.u-tokyo.ac.jp

1 はじめに

近年、計算言語学の世界では、大量のテキストデータ (コーパス) が蓄積されるようになってきたことから、それらのコーパスに対して様々な情報を付与 (アノテーション) し、アノテーションされたコーパスから言語処理用知識を得る手法が一般的に用いられている。それゆえ、コーパスのアノテーションは計算言語学の世界で重要なテーマの 1 つになっている。一般にテキストデータに対するアノテーションとは、テキストデータ中の単語もしくは単語列を指定し、指定された単語・単語列に何らかの情報を付与することである。

人手によるアノテーションにおける問題の一つとして、アノテーションの一貫性の維持の困難さがあげられる。ある情報をテキストに付加するとき、同じ情報をアノテーションするとしても、単語・単語列の領域の指定の仕方にずれが生じたり、どの単語・単語列にアノテーションしたらいいか判断が難しい場合がある。加えて、アノテーション作業は非常に多くの時間を要し、しばしば数週間・数か月かかる。そのため、異なるアノテーター間で一貫性が損なわれる危険性が常につきまとってくる (inter-annotator discrepancy)。それどころか、1 人 1 人のアノテーションを行う人たち (アノテーター) においても、時間の経過につれ一貫性が狂う可能性がある (intra-annotator discrepancy)。アノテーターはどのようにアノテーションしたらいいか悩んだ場合、一貫性を維持するためにアノテーションガイドラインを参照することになる。アノテーションガイドラインとは、アノテーター同士で決めておくアノテーション方針である。そのため、アノテーションガイドラインの作成が重要な課題になる。

アノテーションガイドラインはアノテーションの一貫性を保つための重要な役割を担っているが、アノテーションにおける判断が難しい事態を初めから全てを予測しておくことはできないので、アノテーション作業を行う前の段階に完全なアノテーションガイドラインを用意することはほぼ不可能である。そのためアノテーション作業と共に、アノテーションガイドラインの管理作業を行う必要がある。我々は、アノテーションだけでなく、アノテーションガイドラインの管理も同時に行うシステムを提案する。

提案するシステムは以下のような流れになる。まず一般的に、アノテーションをいかに行うか判断が困難な事例に遭遇し、アノテーションガイドラインを参照しても判断できない場合、アノテーターは自分の直感に従ってアノテーションを行う。しかし常に一意に判断できるわけではないため、アノテーターはどのようにアノテーションするか決断しなければならなくなる。このとき、アノテーションされた結果だけではなく、アノテーション作業中の決断の基準もできているはずである。このようなアノテーション作業中の決断の基準は、アノテーションガイドラインの管理にとって非常に有用だが、アノテーション上の決断の基準を取り扱うアノテーションシステムは存在しない。我々は、アノテーション上の決断の基準を収集したものを用いて、アノテーションガイドラインを管理する手法を考える。アノテーション作業を行いながらアノテーション上の決断の基準を収集するシステムが存在すれば、アノテーション作業とアノテーションガイドラインの管理を別々に行うような余計な手間をかける必要がなくなる。そのうえ、アノテーション上の決断とアノテーションの実例を結びつけて管理することができるため、より有用なアノテーションガイドラインが得られると考えられる。

本論文では、アノテーション作業中にアノテーションガイドラインの管理に必要な情報を収集することができる半自動的アノテーションシステムの提案を行う。Section 2 では、本論文で前提としているオントロジーに基づくアノテーションについての解説、Section 3 では、アノテーションガイドラインの管理手法についての提案、Section 4 では、本手法に基づいたアノテーションツールの作成について説明する。

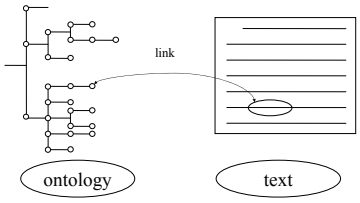

図 1: テキスト上の言語表現とオントロジー上の知識定義とを結びつけるリンク

2 オントロジーに基づくアノテーション

自然言語処理の世界において、アノテーションは計算機に読み込める形式で行われることが多い。一般的には、アノテーションを行うための記述子としてラベルのセットを作成し、そのラベルをテキストの対応箇所に付与することでアノテーションするという形で行われる。とりわけ近年においては、オントロジーを用いたアノテーションが主流なってきた。オントロジーの多くは、計算機で解析できる形式で構築されている。アノテーションで用いるラベルをオントロジー上の知識を用いて定義することで、コーパスに計算機で解析できる情報を、オントロジーで体系的に管理できる形で付加することができる。例えば、『名詞』というラベルがオントロジー上の知識によって定義されているとき、『名詞』のラベルがつけられた単語列は全て、オントロジー上の『名詞』という知識の定義に一致することが言える。

アノテーションを行うことで、テキストはオントロジーによって情報を付加される。同時に、アノテーションされたテキストはそれ自身が知識の実例になり、オントロジー上の知識定義をより深めることになる。例えば、『名詞』のラベルがつけられた単語列は、オントロジー上の『名詞』という知識定義を深める実例になる。このように、テキストとオントロジーとが相互に作用し合うことが、オントロジーを用いたアノテーションの利点といえる。

この観点から、アノテーション作業はテキスト上の言語表現とオントロジー上の知識定義とを結びつけるリンクを結ぶ作業とみなすことができる (図 1)。我々は、このリンクの集合を『コーパス-オントロジー間マップ』と呼ぶことにする。

3 アノテーションガイドラインの管理

本章では、アノテーション作業を行いながら効率的にアノテーションガイドラインの管理を行う手法を提案する。

3.1 管理手法の提案

本手法では、アノテーション作業をテキスト上の言語表現とオントロジー上の知識定義とをリンクで結びつける作業だと考える。一般的に、アノテーション作業を行う前の段階では、アノテーションガイドラインはアノテーションの基本方針を提案したものに過ぎない。そのため、アノテーターがアノテーションをいかに行うか判断が困難な事例に遭遇した場合、アノテーターはアノテーションをいかに行うか決断しなければならない。

このことをリンクの概念を用いて捉えると、アノテーションをいかに行うか判断が困難な事例はリンクの候補だと考えられ、アノテーターが行う決断は候補のリンクが本当に繋がっているか否かの決断だと言える。この決断で得られたリンクの候補と決断の結果・理由は、今後のアノテーションを行う上で非常に有用な情報であるので、これらを用いてアノテーションガイドラインの更新、修正追加などの管理を行う。このような作業を繰り返し行い、アノテーション作業とアノテーションガイドラインの管理を同時に行う。

3.2 アノテーション-メタデータ

Section 2 では、『コーパス-オントロジー間マップ』の定義を行った。ここで、『コーパス-オントロジー間マップ』に付加する情報として『アノテーション-メタデータ』を定義し、アノテーションによって得られたリンクの候補と決断の結果・理由を扱う手法について提案する (図 2)。『アノテーション-メタデータ』は、『コーパス-オントロジー間マップ』の持つテキスト上の言語表現とオントロジー上の知識定義とを結びつけるリンクに対して 3 つの情報を付加する。1 つ目は、リンクの候補が本当に繋がっているか否かの決断 (decision)。2 つ目は、決断を行うために用いる視点 (description)。3 つ目は、決断を行うまでに参考にした他の『アノテーション-メタデータ』(reference)。description では、テキスト上の言語表現とオントロジー上の知識定義との関係性を表す。関係性がはっきりすることで、決断を行うときの指標になる。この指標を残すことで、これからアノテーションを進めて、アノテーションの一貫性の維持が難しい事態に遭遇したとき、『アノテーション-メタデータ』を付加した『コーパス-オントロジー間マップ』を参考にすることができる。参考にされた『コーパス-オントロジー間マップ』は、『アノテーション-メタデータ』として保存された決断の実例になるため、アノテーションの一貫性の維持のために非常に有用な情報になる。このとき、決断の結果を decision に、

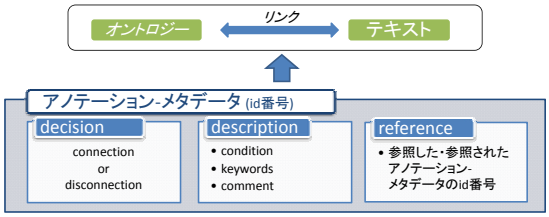

図 2:『コーパス-オントロジー間マップ』と『アノテーションメタデータ』

参考にした・されたという『アノテーション-メタデータ』の関係性を reference に残す。

description はさらに 3 つの部分で構成される。1 つ目は、主にテキスト側の環境の説明 (condition)。2 つ目は、主にオントロジー側の環境の説明 (keywords)。3 つ目は、自然言語で書かれた description の情報 (comment)。condition はアノテーション予定のラベル、アノテーション対象の単語列などの情報が入る。これらには、リンクの候補を探し出す基準として、または検索によってリンクの候補を探し出す場合に用いるクエリとして用いる。condition に対する詳しい説明を、comment に自然言語で追加する。comment は、リンクの候補が繋がっているか否かの決断したときの理由が含まれる。このとき comment に用いた用語を、keywords に登録する。keywords に登録した用語は、オントロジー上の知識情報によって定義する。これにより、テキストとオントロジーとの間の関係性をより明確にすることができる。keywords は、前に作られた description を検索で探し出すときに用いられる。

3.3 アノテーションの流れ

『アノテーション-メタデータ』を用いたアノテーションの流れは、以下のようになる。アノテーターはアノテーションガイドラインに従いアノテーションを行う。アノテーションをいかに行うか判断がこんな事例に遭遇した場合、テキストとオントロジー間のリンクの候補を張り、リンクの候補が繋がっているか否かの決断を行う。このとき、決断の結果を decision に残し、リンクの候補を選んだ理由と決断の理由と用いて description 作成する。アノテーションしている最中に、参考にしたい過去のアノテーションを探し出す場合、現在のアノテーションに類似したアノテーションを行った例を探し出す必要がある。description にはテキストとオントロジーとの間の関係性が明確になっているため、類似したアノテーションを探すための基準になる。また、決断を行う前に過去の『アノテーション-メタデータ』を参考にした場合、その参照先を reference に残す。以上のように、decision・description・reference をまとめて一つの『アノテーション・メタデータ』とし、リンクの候補へ付加する。得られた『アノテーション・メタデータ』を用いてアノテーションガイドラインを更新・追加・修正して管理し、アノテーション作業を再開する。

3.4 アノテーションの例

ここで、実際のアノテーションの例を考えてみよう。Named entity のアノテーションにおいて、プロテインだと思われる単語列に “ $< protein > \sim < /protein >$” というラベルをつけるというアノテーションガイドラインがあるとする。単純に “$protein$” という単語が出てきた場合は、“ $< protein > protein < /protein >$” というラベルを付ければいい。しかし “$NFkappaBprotein$” という単語が出てきた場合 (“$NFkappaB$” はプロテインの種類の名前) は、“ $< protein > NFkappaBprotein < /protein >$” とつけるべきか “$NFkappaB < protein > protein < /protein >$” とつけるべきか 2 つの候補が考えられ、簡単には決断できない。この場合、この 2 つの候補がリンクの候補となる。

過去の『アノテーション-メタデータ』を参照して決断できる場合は、参照した『アノテーション-メタデータ』を reference に保存しておく。そして、アノテーターの決断により、どのようにアノテーションすべきか決める。今回の例では “ $< protein > NFkappaBprotein < /protein >$” とアノテーションするとする。“ $< protein > NFkappaBprotein < /protein >$” のリンクの候補には、リンクが繋がっていることを decision に記述し、その決断の理由を description に記述し、そして reference と共に『アノテーション-メタデータ』に保存し、リンクの候補へ付加する。

対して、“$NFkappaB < protein > protein < /protein >$” のリンクの候補には、リンクが繋がってないことを decision に記述し、その決断の理由を description に記述し、そして reference と共に『アノテーション-メタデータ』に保存し、リンクの候補へ付加する。

その後、2 つのリンク候補に付加された『アノテーション-メタデータ』は、アノテーションガイドラインに追加される。

4 ツール

本章では、Section 3 で提案した手法に基づき、人手によるアノテーションをサポートするツールの提案を行う。このツールでは、『アノテーション-メタデータ』が人手によるアノテーションの一連の手順を保持することができる。人手によるアノテーションの手順には様々な形式がある。本システムは様々な形式に対応で

きるが、説明のためにここでは、各ラベルに対し検索によってアノテーションされる候補を絞り込み、絞り込まれた候補を人手でチェックし実際にアノテーションを行うという手順について考える。この手順では『アノテーション-メタデータ』は、検索で用いたクエリーの情報や、候補のチェックの結果、結果についての説明を保持することになる。

我々のサポートシステムは、人手のアノテーションの手順をモデル化し、モデル化された一部のアノテーション作業を自動化することができる。これにより、アノテーションにかかるコストを削減でき、人手によるアノテーション作業では困難だった、一貫性のあるアノテーションをサポートすることができる。

実際にはどのようなモデル化が行われたかを説明するために、このシステムである単語列にラベル A をアノテーションする手順について説明する。まず、アノテーターはラベル A をアノテーションするために、アノテーション対象となる単語列の候補を検索する。そのあと、検索によって得られたアノテーション候補の単語列を、アノテーションガイドラインを用いて絞り込む。このとき、アノテーションガイドラインに基づき、アノテーション対象となる単語列や周辺の単語などを手がかりにする。実際にラベル A をつけてアノテーションを行う。

このシステムを用いたアノテーションでは、アノテーション作業は 1:アノテーション候補の検索　2:アノテーション候補の絞り込みの、大きく 2 つのステップに分けられ、モデル化される (図 3)。この 2 つのステップのなかで『アノテーション-メタデータ』の作成・修正が行われる。

アノテーション作業は以下の流れになるまず、description の作成を行う。この description はアノテーションガイドラインに基づいて作成される。description には、アノテーション候補の検索に用いるクエリや、アノテーション候補の絞り込みに用いる情報が含まれる。この description を用いて、アノテーションの候補の検索を行う。検索に用いたクエリは、『アノテーション-メタデータ』の description の condition に登録される。次に、アノテーション候補の絞り込みを行う。絞り込みに用いた決断の基準・理由は、description の comment に登録される。このとき、絞り込みによって多くのアノテーション候補が不適切だと判断された場合、アノテーション候補の検索に用いたクエリが不十分だったことが考えられる。また、絞り込みでアノテーション候補が適切か不適切か判断しにくい場合がある。この場合、再度検索・絞り込みというステップを行い、行った結果を用いて『アノテーション-メタデータ』の修正を行う。このように、このシステムを用いたアノテーション作業は、繰り返し作業が行われる。この繰り返しは、クエリのエラーが無くなり (または、十分に少なくなり)、絞り込みが十分に行われたときに止める。

本論文では一例として、Eclipse [1] のプラグインという形としてツールの実装を行った (図 3)。

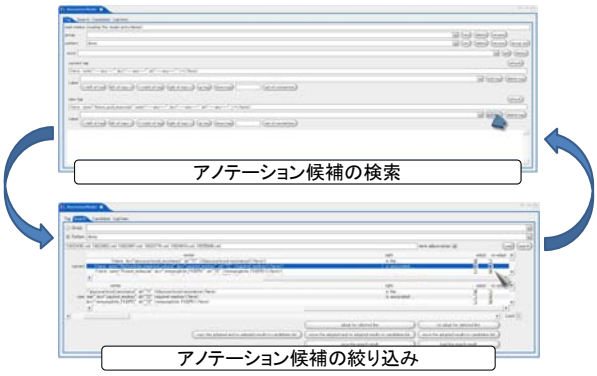

図 3: ツールを用いたアノテーションの手順

5　最後に

本論文では我々は、アノテーションガイドラインを系統だった手法で管理する方法として、アノテーションシステムを提案した。アノテーションガイドラインは十分な実例とセットで管理することで、実例と似た事例において、一貫性を保ったアノテーションを行うことができる。アノテーションガイドラインと実例を結びつけるために、我々は『アノテーション-メタデータ』を提案し、『アノテーション-メタデータ』を用いたツールを実装した。このツールを用いることで、人手によるアノテーション作業の一部 (たとえば、アノテーションガイドラインを管理し参照する作業) を自動化し、アノテーションによるコストを軽減し、同時にアノテーションの質を上げることができる。今回、熟練のアノテーターの方々にアドバイスを頂き、システムを提案した。

今後の予定として、ツールを実際に使用しアノテーションにかかるコストがどれくらい軽減されるか、どれくらいアノテーションの一貫性が保てるか評価を行いたい。

参考文献

[1] Eclipse - an open development platform. http://www.eclipse.org/.