照応関係を示したタグ付き英語コーパスの作成

若尾 孝博　通信・放送機構 (TAO)
江原 暉将　ＮＨＫ放送技術研究所 / 通信・放送機構
白井 克彦　早稲田大学 / 通信・放送機構

1　はじめに

１９９５年の秋に米国で開催された第６回 Message Understanding Conference (略称 MUC-6、米国国防省 Defense Advanced Research Projects Agency (DARPA) 主催) において、情報抽出作業の１つとして英語新聞記事中の照応関係 (coreference relations) を自動的に認定する作業があった ([1], [2], [3])。

MUC-6 では照応関係の自動認定作業の評価をするために、正解となる照応関係を明示したテキスト (英語ビジネス新聞記事) を人手をかけて用意した。

ここでは、テキスト中の照応関係を示したコーパスを如何に作成したかを、「何を照応関係とみなすか」、それを「テキスト中でそれをどう明示するか」を中心に、作成の過程を追って報告する。[1]

2　コーパス作成参加者と期間

照応関係を示したコーパスを作成したのは、あくまでボランティアの研究者であり、その全員が過去に行われた MUC の参加経験者で、１２の大学または企業からの研究者が参加した。コーパス作成は１９９４年６月頃から始まり、１年３ヵ月後の１９９５年９月に最終版が完成した。テキストは The Wall Street Journal を用いた。取りまとめ役は、米国政府側では Ms. Beth Sundheim であり、研究者側は Professor Ralph Grishman (New York University) であった。議論の媒体は専ら電子メールでのやりとりが中心であった。

[1]若尾は当時英国の University of Sheffield に居り、このコーパスの作成に参加した。

3　Hobbs 氏の案と作業仕様第１版

コーパスを作成するにあたって、まず「どのような照応関係を作業の対象とするか」が問題となった。これに対して、まず SRI の Dr. Jerry Hobbs が次のような作業仕様を提案した。これは１９９４年６月に提案され、その後議論の出発点となった。

1. 文中の各単語に番号を付ける。固有名詞や "set up", "joint venture" などの複合語はハイフンで結び一語として扱う。
2. テキスト中において照応関係の認められる全ての固有名詞、代名詞、名詞句について、その表現 X (the refering expression X) とその係り先 Y (the antecedent Y) 及び X Y 間の関係を特定する。いろいろな関係が認められるが、下記の５種類に絞る。
   - (Ident X Y):
     X と Y は同等 (identical) である
   - (Sub X Y):
     X は Y のサブセット又は１要素である
   - (Sub Y X):
     Y は X のサブセット又は１要素である
   - (l-subj Y X):
     Y は X を名詞化したものの論理主語
   - (l-obj Y X):
     Y は X を名詞化したものの論理目的語
3. 更に、これ以外の関係を示したい場合には (Rel X Y) とする。
4. 番号付けにあたっては、i.j (i 番目の文の j 個目の単語) を用いている。
5. 常に初出の表現が参照されるものとする。

例文 1

```
1. BRIDGESTONE-SPORTS-CO. 2. SAID 3. FRIDAY
4. IT 5. HAS 6. SET-UP 7. A 8. JOINT-VENTURE
9. IN 10. TAIWAN 11. WITH 12. A 13. LOCAL
14. CONCERN 15. AND 16. A 17. JAPANESE
18. TRADING-HOUSE 19. TO 20. PRODUCE
21. GOLF-CLUBS 22. TO 23. BE 24. SHIPPED
25. TO 26. JAPAN.
```

上記の例文1で認められる照応関係は (Ident 1.4 1.1) だけである。つまり IT (1.4) と BRIDGESTONE-SPORTS-CO. (1.1) が同等関係 (Ident) にある。

この Hobbs 氏の案が出された後、９４年９月末に政府側から作業仕様 (task definition) の第1版が示された。基本的な内容は以下の通りである。

1. Hobbs 氏とは違い、テキスト中に SGML のタグを挿入して照応関係を示す。これは最後までこの形式が残った。例えば、次のようである。

   ```
   <COREF ID="100"> Lawson Mardon Group
   Ltd. </COREF> said <COREF ID="101"
   TYPE="IDENT" REF="100"> it </COREF> ...
   ```

   ここでは、REF="100" が照応先の ID 番号を示している。つまり、it が会社名 Lawson Mardon Group Ltd. を指し示し、照応関係があることを現している。
2. ５種類の照応関係を作業の対象とした。
   - IDENT for "identical," 同等関係
   - PT-WH for "part/whole," 部分と全体
   - WH-PT for "whole/part," 全体と部分
   - SUB-SUP for "subset/superset," 下位セットと上位セット
   - SUP-SUB for "superset/subset." 上位セットと下位セット
3. タグ付けの対象は次の通り。
   - 代名詞、所有格代名詞
   - 限定名詞句 ("the" はじまる名詞句)
   - 固有名詞
   - 特定の副詞句 (例："there", "then")
4. 同格 (apposition)、部分を示す語句 (partitive)、所有格を示す語句 (genitive) や、関係詞節 (relative clause) は対象としない。

4 問題点

作業仕様第1版に基づき、１２の大学、企業のグループが実際の新聞記事 (数十記事) に照応関係のタグ付けを行った。その結果作業仕様そのものに問題があることが分った。次に、第1版作業仕様で問題となったもののうち５つを選び紹介する。

4.1 Sub-Sup, Part-Whole 関係

認定すべき照応関係である、Sub-Sup 関係や Part-Whole 関係については、詳しく定義することが難しく、具体例をあたってみると簡単ではない。例えば、一文中に 'Nissan' と 'Japan' が現れ、これらを Part-Whole の関係とみなした人があった。また

(1) Ford announced a new product line yesterday. Ford spokesman John Smith said ...

という文の場合、'Ford' とスポークスマンである 'John Smith' 氏の関係は果して Part-Whole 関係といえるのかなどの意見がまとまらなかった。つまり、どこまでのは範囲を Part-Whole 関係の対象とするのかが明確にするのが困難であった。また 'furniture' などの集合名詞について、'chair' という語が出て来た場合に、これは Sub-Sup とするのか、Part-Whole とするのかはっきりしないなどの問題があった。

4.2 Full NP 対 NP Head

対象となる名詞句のどこまでを、つまり、名詞句の head だけをマークするのか、それとも名詞句全体をマークするのかで意見が分れた。名詞句の head をマークするのが良いとする意見が当初優勢であった。名詞句の head を正確にマークしたコーパスはなく、この作業をすることにより、

副産物として名詞句の head を明示したコーパスが出来ることになり意義があるとの意見もあった。しかし、結局、これを行うには、まず、「名詞句の head」とは何かを正確に定義する必要があり、しかも、それを見つけるには、名詞句の構造解析をする必要が生じ、本来の照応関係認定の作業に加えて新たな負担がかかるという理由から、結局名詞句は全体をマークすることにして、head を示すには SGML のタグ内に MIN と言う値を設けて対処することで意見がまとまった。

4.3 接続詞で結ばれた名詞句

(2) The boys and girls enjoy their breakfast.

文 (2) のように名詞句が 2 つ以上接続詞で結ばれている場合が問題となった。議論の結果、2 つ以上の head を含む名詞句は照応関係認定の対象から外すことになった。

(3) Fred Smith and Harry Edmond both had dad's apple pie. Harry survived.

しかし、問題はまだ残っており、(3) の文では接続詞で結ばれた名詞句の一部 ('Harry Edmond') が後の文で照応されている。このような場合は照応関係を認めることになった。つまり、'Harry Edmond' と 'Harry' の間に IDENT (同等) の照応関係があると認めることにした。

4.4 換喩 (Metonymy)

比喩表現の一種である、換喩表現についてどう対処するのかも議論された。例としては以下のようなものがある。

(4) *Ford* announced a new product line yesterday. ... *They* will start manufacturing widgets.

(5) I bought *the New York Times* this morning. I read that the editor of *the New York Times* is resigning.

(4) において、'Ford' は会社名であるが、'they' と照応している。(5) ではじめの 'the New York Times' は新聞そのもの自体を言っているのに対して 2 番目の 'the New York Times' は会社組織を指している。これらの換喩表現に関しては、しっかりとしたガイドラインが示されず、結局マークをしてもしなくても良い「オプション」とすることになった。

4.5 タグ付けの結果の一貫性

第 1 版の作業仕様に基づいてタグ付を付けられた記事の幾つかは 2 つ以上のグループで重複して、しかし、互いに独立してタグ付けがなされた。その結果についてどれだけ一貫性があるかを調べてみると、recall が 5 9 %、precision が 7 1 %と、一貫性が見られず、作業を行う人の判断に頼るところが多く、個人差が現れた結果となった。これは大きな問題と認識され、どうすれば高い一貫性が得られるのかが議論された。その結果、一貫してタグを付けを行おうとする際に一番問題だとされた「Sub-Sup, Part-Whole 関係」を作業対象から外した。つまり、対象となる照応関係は IDENT とのみとすることに決定した。これで一貫性が増し、目標であった recall と precision において 8 0 %以上を確保することが可能となった。

5 作業仕様最終版

実際の新聞記事に照応関係を付ける作業を行い、作業仕様を改良していった。最終の作業仕様は、1 9 9 5 年 9 月に作成された。版はこの時 2 . 3 版となっていた。概要は以下の通りである。

1. 第 1 版と同じ形式の、SGML のタグを採用。
2. 対象となる照応関係は IDENT 1 種類のみとする。
3. タグ付けの対象は、基本的に名詞、名詞句、代名詞である。
   - 固有名詞は対象とし、複数語で構成されてる名前は 1 つの分割出来ないものとして取り扱う。
   - 動名詞は対象としない。
   - 代名詞にはその所有格を含み、人称代名詞も対象とする。
   - 明示されない代名詞 (zero pronoun) は対象外とする。例えば、

(6) Bill caleed John and spoke with him for an hour.

文 (6) において 'Bill' は 'spoke' の主語であるが、この2つの語句の関係を認定することは作業の範囲ではない。

4. 名詞句の範囲は、対象と思われる名詞句全体をマークする。同時に出来ればその名詞句の head をタグの中に示すものとするが、これはオプショナルである。
5. 接続詞で結ばれた名詞句
   前記の議論どおり、2つ以上の head を含む名詞句は照応関係認定の対象から外す。しかし、その一部が後出の名詞句によって照応されるばあいは、照応関係を認めるものとする。
6. 換喩表現 (Metonymy)
   オプショナルとする。
7. 同格 (apposition)
   対象とした英語新聞記事には、頻繁に

(7) Julius Caesar, the well-known emperor,

などの同格表現が現れた。このような同格表現では、第1の語句が第2の語句によって説明されている。この場合、照応関係の明示の仕方は、第2の語句が表現全体を指し示しているとする。文 (7) では、'the well-known emperor' と、この同格表現全部、'Julius Caesar, the well-known emperor,' の間に照応関係があると認めることになる。

この最終版に基づいて照応関係のタグ付けをしたコーパスが、95年秋に開催された MUC-6 大会における評価に用いられた (作業仕様 (task definition) の詳しくは英語で [1] に掲載されている)。この作業仕様で行った場合のタグ付け結果の一貫性は、平均で recall 80%、precision 82%となった。

尚、本報告では一切触れなかったが、照応関係自動認定の作業を評価するにあたり、人の手によってタグ付けされたテキストと、機械によって作成された照応関係のタグ付きテキストをどのように比較し、採点するかは、別の問題としてかなりの議論を呼んだ。特に、同等関係の長い鎖があるとき、システムがその一部だけを認定した場合どうするかなどが問題となった。これについても MUC-6 の予稿集 ([1]) を参考にして頂きたい。

6 最後に

最後のなったが、MUC-6 大会での Coreference Task (照応関係の自動認定作業) の成果について触れておく。この作業には全部で7個のシステムが参加した。大半のシステムが同レベルの成績を残した。上記のようにして人手で作成した正解と比べてみて、7個のシステム中5個が recall で51%～63%、precision で62%～72%の範囲に収まる成績であった。米国国防省が支援する MUC 大会は今後も続く予定で、1997年に MUC-7 が開催される予定である。照応関係の自動認定作業は、MUC-7 においても引続き含まれると思われ、タグ付きのコーパスのサイズもより大きなものになって行く様子である。

本報告では、米国で行われた英語新聞記事を用いての照応関係を示したタグ付けコーパスの作成の過程について、概略を述べた。1994年6月頃から始まり、1年3ヵ月の間には、ここに書かれていない多くのことが討論され、議論された。ここでは紙面の制限上省いているものもあることをご了解頂きたい。今後、もし日本において、日本語テキストを対象とした照応関係のタグ付けコーパスが作成されることがあり、その際の参考になれば幸甚である。

参考文献

[1] *Proceedings of the sixth Message Understanding Conference (MUC-6)*, Columbia, Maryland, U.S.A. 1995, Morgan Kaufman Publishers Inc. 1996.

[2] Ralph Grishman and Beth Sundheim "Message Understanding Conference - 6: A Brief History" In *Proceedings of the 16th International Conference on Computational Linguistics (Coling 96)*, 1996.

[3] 若尾 孝博, "英語テキストからの情報抽出：MUC 第6回大会の参加報告", 電子情報通信学会技術研究報告 NLC-96-9-20, 1996.