単一言語のアラインメント

飯伏勝俊　鳥澤健太郎　辻井潤一
東京大学理学部情報科学科
{k-ibushi, torisawa, tsujii}@is.s.u-tokyo.ac.jp

1. はじめに

同一の事件、事象を自然言語によって記述した場合、その表現は必ずしも一意ではない。このことは、計算機による知識獲得や検索の妨げとなることが多々あり、自然言語の表現の多様性を捨象する方法が必要となる。

本研究では、日本語で記述されている日本経済新聞と毎日新聞の記事コーパスから同一の事件を記述している記事を多言語間アラインメントで用いられている手法を応用して対応づけた後、同じ手法を用いて、同一の事象を記述している文を対応づけている。

多言語間アラインメントの場合は、英語で記述された文章とその邦訳のように文章・文の構造が似通ったコーパスが用いられる。しかし、本研究では同じ事件・事象を記述していても、文章・文の構造が異なるコーパスを用いている。このため、多言語間アラインメントで用いられるベクトルモデル[3]をそのまま適用した場合、アラインメントを正確に行うことが困難となってしまう。そこで、ベクトルモデルに制約を加えることによって、アラインメントの正確さを実現している。

2. アラインメントの手法

本研究では、単一言語で記述された2つの新聞記事データ間での記事・文アラインメントに多言語間アラインメントで用いられているベクトルモデルを応用する。記事・文のアラインメントには、ほぼ同一の手法を用いているので、以下では記事のアラインメントの手法について説明し、文の場合との差異は本章の最後で説明する。

用いた手法は、形態素ごとにその出現頻度によって重みづけを行い、記事間の類似度を記事間で共起した形態素の重みと出現数から計算して評価するというものである。

2.1. 重みづけ

頻度の低い形態素ほど重みが大きくなるように設定する。これは、頻度の低い形態素は、2つの記事に偶然に共起する可能性は低いので、形態素が共起した記事の間には強い関連性が存在すると考えられ、逆に、頻度の高い形態素は複数の記事において共起する可能性が高いので記事同士の関連性を示す目安にはなりにくいためである。

idf [1]は多言語間の文章の対応づけや、関連記事の検索、カテゴリ分類に用いられている代表的な重みづけの手法である。形態素 t の重みは以下のように定義される。

$$idf(t) = \log(N/df(t))$$

ただし、N はアラインメントを試みる記事の総数であり、$df(t)$はそれらの記事のうち形態素 t を含んでいるものの総数である。

2.2. 記事間の類似度

多言語間アラインメントは主にある言語で記述した記事と、その記事を他の言語に翻訳した記事との間で行われる。この場合、記事 d を以下のベクトルで表現するベクトルモデ

ルが主に利用されている[3]。

$$W_d = \left(w(d,t_1),\cdots,w(d,t_n)\right)$$

$$w(d,t_i) = tf(d,t_i)\cdot idf(t_i)$$

ただし、$tf(d,t_i)$は記事 d 中での形態素 t_i の出現回数である。そして、記事 q, dの類似度をこのベクトルの内積で以下のように表す。

$$I(q,d) = \sum_i tf(q,t_i)\cdot tf(d,t_i)\cdot (idf(t_i))^2$$

しかし、単一言語アラインメントの場合、この類似度 $I(q,d)$ではよい結果を得られない。これは、$I(q,d)$が以下の仮定の下で、類似度とされているからである。

- 同一の記事を表現するベクトルは平行になる
- 平行なベクトル間の内積は大きくなる

今回の実験のように文章の構造が異なる記事同士を対応づけようとする場合、同一の事件を記述していても、それぞれの記事を表現するベクトルは平行であるとは限らない。

また、記事 dに偶然1回出現した形態素であっても記事 q に多数出現していれば、類似度を大きくすることになる。また、この逆もいえる。

そこで、ベクトルモデルに次のような制約を加える。

- 日経の各記事をクエリー qとし、以下のベクトルで記事を表現する。

$$Q = (w_{q1},\cdots,w_{qn})$$

$$W_{qi} = \begin{cases} 1\left(t_i \in q\right) \\ 0\left(t_i \notin q\right) \end{cases}$$

- クエリー q と毎日の各記事 d との類似度を以下のように表現する。

$$S(q,d) = \sum_i w_{qi}\cdot tf(d,t_i)\cdot idf(t_i)$$

$I(q,d)$と異なり、$S(q,d)$は記事 q, d 間における形態素の出現頻度の類似性は考慮されていない。これは、以下の仮定のに基づいている。

- 記事 dを特徴づけているのは、記事 d 中に多数出現する形態素である
- 記事 d を特徴づける形態素が記事 q 中に出現しているなら、記事 q, d の間には関連性がある。

また、$S(q,d)$は $I(q,d)$の「記事 d に偶然1回出現した形態素であっても記事 q に多数出現していれば、類似度を大きくする」という欠点を改良している。ただし、逆の場合は考慮されておらず、また $S(a,b)$と $S(b,a)$が違う値となるという欠点が存在する。

2.3. 形態素出現数の正規化

一般に長い記事の場合、関連性のない記事同士であっても形態素の共起が起こる可能性が大きくなる。したがって、これまで定義した $I(q,d)$と $S(q,d)$は長い記事ほど大きくなる傾向がある。

そこで、$tf(d,t_i)$を記事の長さ l_d によって正規化し、$NI(q,d)$と $NS(q,d)$を以下のように定義する。

$$l_d = \sum_i tf(d,t_i)$$

$$NI(q,d) = I(q,d)/(l_q\cdot l_d)$$

$$NS(q,d) = S(q,d)/l_d$$

本研究では、以上のように定義した $I(q,d)$, $S(q,d)$, $NI(q,d)$, $NS(q,d)$の4つの類似度を比較した。

2.4. 記事と文のアラインメントの相違

本研究では新聞記事を日本語形態素解析システム JUMAN[2]で解析し、「形態素　品

詞」という形式で区別し、各記事ごとの各形態素の出現回数を数えた。

記事のアラインメントにおいては、「普通名詞」「人名」「組織名」「地名」「サ変名詞」「時相名詞」といった名詞に属する形態素を用いた。これは、記事を特徴づける品詞として、名詞が他の品詞よりも有効であるからである。

しかし、文の場合は名詞だけでは文中での出現数が少ないので、その文を特徴づけるのには不十分である。そこで、文のアラインメントの場合は名詞に加えて動詞、数詞に属する形態素の出現回数も利用して類似度を計算する。

また、idf はアラインメントの対象となる文章中での文の総数を N とし、N の中で形態素 t_i を含む文の数を $df(t_i)$ とする。

3. 実験

Terminology

4つの類似度の評価を行うために以下の2つの指標を導入する。

$$\text{適合率} = \frac{\text{システムが対応づけたペアの中で正解のペアの数}}{\text{システムが対応づけたペアの数}}$$

$$\text{再現率} = \frac{\text{システムが対応づけたペアの中で正解のペアの数}}{\text{正解のペアの数}}$$

3.1. 記事アラインメントの結果

記事のアラインメントを、1994 年 4 月 4 日の日経 314 記事、毎日 258 記事について行った。これらの間には対応する記事が 61 組存在する。

実験の結果は図 3.1 に示す。$NS(q,d)$ を類似度に用いたものが再現率 0.5 以下の範囲では適合率 0.8 以上となっており、最良の結果を示している。しかし、再現率が 0.5 を越えると急激に適合率が落ちてしまっている。

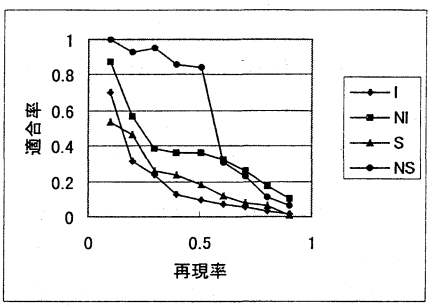

図 3.1: 記事のアラインメントの結果

再現率 0.5 から 0.6 の範囲で誤ってアラインメントされているのは、下記のような記事同士であった。

a) NGO についての記事と郵政省の国際ボランティア貯金についての投書
NGO の活動資金として国際ボランティア貯金が触れられていた。

b) 日経の景気動向の観測記事と物価問題に関する世論調査についての記事
後者で、コメが値上がりしたと感じた人が多かったことが繰り返し強調されていたが、前者の文中でコメという単語が1回出現していた。

これらは、$S(q,d)$, $NS(q,d)$ の欠点である「記事 q に偶然1回出現した形態素であっても記事 d に多数出現していれば、類似度を大きくする」に合致して失敗した例である。

3.2. 文のアラインメントの結果

文のアラインメントでは、4月4日の記事のアラインメントで再現率 0.5 の際に対応づけられた 35 記事を実験に用いた。ただし、4記事は誤ってアラインメントされている。

実験の結果は図 3.2 に示してあるが、文のアラインメントにおける各類似度の差は記事のアラインメントの場合ほど大きくはない。しかし、$NS(q,d)$ を用いた場合、100 組の文

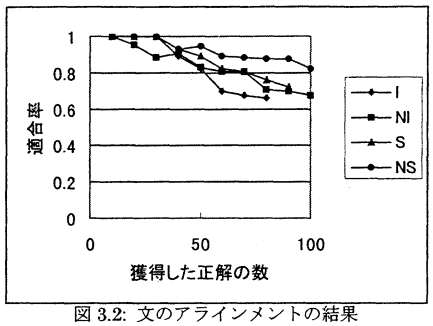

図 3.2: 文のアラインメントの結果

を獲得しても適合率が 80%を割っていない。

また、記事のアラインメントの誤りが文のアラインメントに及ぼした影響としては、日韓外相会談の記事と細川首相・韓国外相会談の記事の間で会談内容が一致するために何について会談したのかという文がアラインメントされていたことなどがあった。

3.3. 考察

今回の実験で、単一言語のアラインメントの場合は、ベクトルモデルをそのまま用いるよりは、制約を加えた方がうまくいくことが確認できた。

ただし、$NS(q,d)$については、事件記事は比較的正確にアラインメントできるものの、政治・経済記事などでカテゴリの近い記事同士をアラインメントしてしまったり、コラム記事のような特殊な記事によって適合率が減少するなどの問題があった。

4. おわりに

本研究で用いた $NS(q,d)$は、再現率が低くなってしまうが、高い適合率で同一の事象を記述している文のペアを獲得できる。このことから、$NS(q,d)$は、数年分の新聞記事を入力として再現できる記事・文の少なさをカバーすることにより、正確さの要求される学習のためのデータを獲得するのに利用できると考えられる。

今後は、獲得できたデータを統計的手法や素性構造を用いて解析し、句や語の多様性についての知識を獲得することが課題となる。多様性についての知識が獲得できれば、それをアラインメントに還元することによって、より多くの文のペアを獲得することができるようになり、更なる知識の獲得が期待できる。

謝辞

本研究で利用したコーパスは、日本経済新聞 CD-ROM '94 版および CD-毎日新聞 '94 版から得られています。コーパスの利用を許可していただいた両新聞社、及び、このコーパスの利用に関して尽力された方々に深く感謝します。

参考文献

[1] K. Sparck Jones : "A statistical interpretation of term specificity and its application in retrieval", Journal of Documentation, Vol. 28, No. 1, pp.11-21(1972)

[2] 松本裕治、黒崎禎夫、宇津呂武仁、妙木裕、長尾眞: "日本語形態素解析システム JUMAN 使用説明書 version3.4", 京都大学工学部 長男研究室, 奈良先端科学技術大学院大学 松本研究室(1997)

[3] N. Collier, 熊野明, 平川秀樹: "多言語情報検索技術を用いた二か国語コーパスの自動アラインメント", 電子情報通信学会, NLP97, pp.39-46(1997)