5Q-01

クラウドソーシングによる確率的アノテーションを利用した談話関係解析

澤田 晋之介 (a　小浜 翔太郎 (b　河原 大輔 (b　黒橋 禎夫 (b(c

(a 京都大学工学部　(b 京都大学大学院情報学研究科　(c JST, CREST

1 はじめに

言語リソース構築のためのアノテーションは, 専門家の手で正解ラベルを付与することによって行われてきた. この方法では高品質なリソースが構築できる一方で, 規模に限界があり, また構築にかかる時間, 金額ともに大きなものとなる. そこで近年, クラウドソーシングを用いて言語リソースを構築する手法が提案されている. この方法なら, 大規模なリソースを短時間で安価に作成することができる一方で, 専門家によるものと比較して品質の劣るリソースになると言われている. この問題に対処するために, 品質管理の手法が提案されている. これらの手法では, 複数人のワーカーによって付与されたラベルを利用して, 正解ラベルに対する確率が計算される [4].

言語リソースの利用方法として, 教師あり学習を用いて解析器を構築することが挙げられる. 従来の手法では, このようなクラウドソーシングによって作成されたリソースを用いる場合であっても, 付与された確率をもとに正解ラベルを決めて専門家が構築したリソースと同じように利用していた.

本研究では, 品質管理の手法によって計算された確率を, 回帰を用いて推定し, それを機械学習で利用する手法を提案する. 河原らが作成した談話関係タグ付きコーパス [5] を用いて実験を行い, 確率を用いない場合と比較して精度が向上することを確認した.

ここで対象とする談話関係解析の先行研究としては, Lin らによる, Penn Discource Treebank[2] を用いたもの [1] などがある. Penn Discource Treebank は長い年月をかけて人手で構築された談話関係タグ付きコーパスであり, クラウドソーシングによって構築されたものではない. また, 関連研究としてクラウドソーシングによるリソースの構築を行っているものが挙げられる. このような研究としては, クラウドソーシングによるリソース作成の先駆けとなった Snow らによる簡単なタスク群 [3], また日本語では今回利用した河原らによる 2 段階のクラウドソーシングによる談話関係タグ付きコーパスの作成 [5] などがある. 品質管理の手法によって計算された確率を, 回帰を用いて推定し, 学習に利用した先行研究はこれまでのところ存在しない.

Discourse Parsing based on Probabilistic Annotations via Crowdsourcing
Sawada Shinnosuke(a, Kohama Shotaro(b, Kawahara Daisuke(b and Kurohashi Sadao(b(c

2 談話関係タグ付きコーパス

本研究で用いたコーパスは河原らが作成したものである. 談話関係有無判定, および談話関係ラベル判定の 2 段階のクラウドソーシングによって作成されている. 3 文からなる文書を 5 節を上限として節に区切り, 節ペアに対して談話関係をアノテーションしている. クラウドソーシングのそれぞれの結果に対して, Whitehill らが提唱した手法 [4] を用いて正解ラベルとその確率を計算している. 談話関係有無判定にて談話関係のある確率が 0.01 を越えた節ペアについて, 第二段階の談話関係ラベル判定にて, 原因・理由, 条件, 目的, その他根拠, 対比, 譲歩, その他弱い関係の 7 つに分類している. 文書数は合計 10,000 であり, 1 文書あたり 3 文であるので文数は合計 30,000 である. 第一段階で談話関係有無確率が 0.01 を下回った節ペア, 及び第二段階でその他弱い関係と判定された節ペアを「関係無し」として扱った. 関係無しの節ペアが 54,105 個, 関係ありの節ペアが 4,024 個であった.

3 提案手法

まず, 確率の付与された節ペアを訓練データとして, 回帰を用いて確率を推定する. 次に, 付与した推定値を素性に加えて談話関係解析器を訓練し, テストセットに対して二値分類を行う. 談話関係解析器に用いた素性を表 1 に示す. これらは, 河原らがコーパス構築時に作成した談話関係解析器のものと同一である.

4 実験

実験は談話関係有無を判定する二値分類タスクを行った. 回帰には SVR, 談話関係有無判定には SVM を用

Copyright ©2015 Information Processing Society of Japan.
All Rights Reserved.

表 2: 確率的アノテーションの利用による改善例

コーパス中の正解ラベル	コーパス中の確率	SVR で推定した確率	文書
関係無し	0.0005	0.1361	一つ残念だったのが２６１１１と２６１４１には有った背面にＡ４サイズの書類が入るようなポケットがあったのに、この２６５１４にはないことです。【これが本当に残念で】【星を一個落とさせてもらいました。】それ以外では満足してます！
関係あり	0.7635	0.3356	食用として、主に煮豆とする。正月料理（おせち料理）には欠かせないものとされる。労苦をいとわず【物事にはげむこと、また、そのさまや、からだのじょうぶなことを「まめ」と呼ぶことから、】【これからの一年をそのように過ごせるようにという験担ぎである。】
関係あり	0.1595	0.1618	胸もとのラメもポイントです！【前ファスナーだから】【サッとはおれるのがいいですね。】長く着られて便利です。

表 1: 談話関係解析器の素性

名前	説明
節距離	節間の距離（節単位）
文距離	節間の距離 （文単位）
単語	節に含まれる単語（原形）
述語	節の述語 （原形）
述語活用形	節の述語の活用形
接続詞	節の最初の単語が接続詞ならその単語（原形）
単語重複率	二節間における単語重複率
節タイプ	KNP によって出力される節のタイプ（約 100 種類）
「は」有無	節中に係助詞「は」があるかどうか
「は」共起	二節ともに係助詞「は」があるかどうか

いた. 回帰, 機械学習のツールとしては SVM^{light} を用いた. カーネルは回帰, 二値分類いずれにおいても二次の多項式カーネルを使用した. SVM^{light} における SVR の C パラメタについてはいくつか試して精度の良かった 0.01 に, SVM の cost-factor パラメータについてはコーパスの分布に偏りがあるため 3 に, 他のパラメータはそれぞれのツールのデフォルトとした. SVR および SVM の素性を抽出するために, 日本語形態素解析器 JUMAN と日本語構文・格解析器 KNP を使用した. 提案手法に対する比較手法としては, SVM 単体による二値分類を採用した. 本研究で用いたコーパスは 2 節で述べたものである. 10,000 文書のうち, 8,000 文書を訓練に, 2,000 文書をテストに用いた. 訓練データへの SVR での推定値の付与は 5-fold jacknifing で行った.

談話関係有無判定タスクを評価するための評価尺度としては, SVM^{light} の出力する決定関数値に閾値を設け, 絶対値が閾値以上のもののみを出力したと仮定して閾値を変化させながら再現率—適合率グラフを描いた. 精度が高いほど右上の方に点が現れる. 図 1 に談話関係有無判定タスクの精度を, 表 2 に改善例を示す.

コーパス中の確率値の平均が 0.048 であったのに対して, SVR で推定した確率値の平均は 0.059 であった. また, その分布は -0.4 から 0.7 の間であった. 表 2 の二つ目の例は SVR での推定値が 0.34 ほどと高めだったことが有効に作用したと考えられる. また, 三つ目の例も推定した確率が 0.16 と高めの値であり, これが有効に作用したと考えられる.

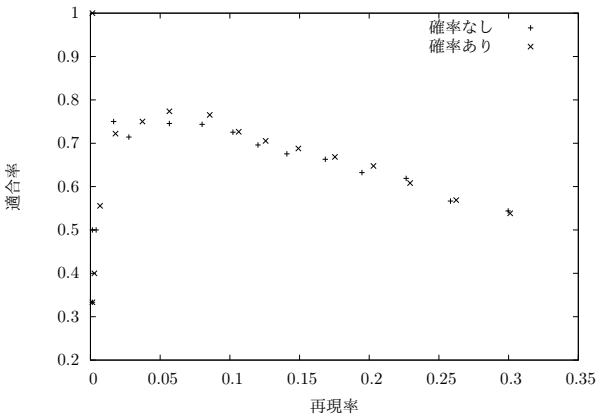


図 1: 談話関係有無判定タスクの精度

5 おわりに

本研究では, クラウドソーシングによって構築されたコーパスを利用する際に, 品質管理の手法で付与された確率を利用することで談話関係解析の精度が向上することを示した. 今後はクラウドソーシングによって構築された, 談話関係以外のタグ付きコーパスに同様の手法を適用し, 有効性を確認する予定である.

参考文献

[1] Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, Vol. 20, pp. 151–184, 2014.

[2] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse Treebank 2.0. In *Proceedings of LREC2008*, pp. 2961–2968, 2008.

[3] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP2008*, pp. 254–263, 2008.

[4] Jacob Whitehill, Paul Ruvolo, Ting-fan Wu, Jacob Bergsma, and Javier Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Proceedings of NIPS22*, pp. 2035–2043, 2009.

[5] 河原大輔, 町田雄一郎, 柴田知秀, 黒橋禎夫, 小林隼人, 颯々野学. 2 段階のクラウドソーシングによる談話関係タグ付きコーパスの構築. 情報処理学会 第 217 回自然言語処理研究会, 2014.

Copyright ©2015 Information Processing Society of Japan.
All Rights Reserved.