現場発想による自然言語処理ブレークスルーの探求

坪井 祐太[1]　森 信介[2]　鹿島 久嗣[3]　小田 裕樹　松本 裕治[4]

[1] 日本アイ・ビー・エム（株）　[2] 京都大学　[3] 東京大学　[4] 奈良先端科学技術大学院大学

〔受賞論文〕
日本語単語分割の分野適応のための部分的アノテーションを用いた条件付き確率場の学習
坪井祐太（日本アイ・ビー・エム（株）），森信介（京都大学），鹿島久嗣（東京大学），小田裕樹，松本裕治（奈良先端科学技術大学院大学）
情報処理学会論文誌，Vol.50, No.6, pp.1622-1635 (2009)

このたび，論文賞をいただき大変光栄である．第一著者は，これまで多数のテキスト分析システム構築に従事してきた．お客様がシステム導入時に感じる不安の多くは，辞書・訓練データ作成などのチューニング作業についてであった．本論文は，訓練データ作成工数を低減するための方法と近年発展した構造データを扱う機械学習手法との乖離解消を動機として始めた研究である．文の中で重要な部分にのみ正解を付与することにより，新しい分野での表現に対応するための訓練データを少ない作業量で作成できる☆1．そこで，正解が付与されていない部分があっても構造出力を学習できる手法を開発し，本論文にまとめた．ここで扱った問題も含め，応用現場は研究課題の宝庫である．以降では，「あいまい性」をキーワードに，より大きな課題について論じたい．

第1は，データを用いて解きたい問題のあいまい性である．本論文での問題は，正解がない部分は正解候補が複数存在する，あいまい性のある訓練データからの学習問題として一般化できる．類似の問題設定としては，人手で正解をひとつに定められないデータからの学習[1]や，正解が範囲として与えられた回帰[2]などを提案してきた．特に，実際の応用では問題定義の過程であっても並行して正解付与作業を進めることがあり，結果としてあいまいな訓練データが作成されることがある．また，訓練データ作成中に各事例を見ることで問題定義が明確になることもしばしばある．今後は問題定義と訓練データのあいまい性を反復的に減少させるための包括的な枠組みの確立に取り組みたい．

第2は，予測のあいまい性である．学習器の予測は誤る可能性があり，ひとつに絞らずあいまいな出力として扱うことで頑健性が向上する[3),4)]．さらに，複数の予測器を連結する際にも予測のあいまい性を考慮することは有用であろう．特に自然言語処理では，複数の予測器を逐次的に繋げたシステムが一般的であり，前段の誤差がシステム全体に伝播することが以前から問題視されてきた．予測出力をひとつに固定してしまうことで後段での修正が困難になることが一因である．訓練データおよび予測出力にあいまい性を許す学習方法の発展が，このような逐次的システムの頑健性向上に繋がると考えている．

最後に，本研究を進めるにあたり有益な議論をしてくださった関係各位，本論文に関して貴重なご指摘をくださった査読者の皆様に心より感謝申し上げたい．

参考文献

1) Tsuboi, Y., Kashima, H., Mori, S., Oda, H. and Matsumoto, Y. : Training Conditional Random Fields Using Incomplete Annotations, International Conference on Computational Linguistics (2008).
2) Kashima, H., Yamasaki, K., Saigo, H. and Inokuchi, A. : Regression with Interval Output Values, International Conference on Pattern Recognition (2008).
3) 岡野原大輔，工藤　拓，森　信介：形態素周辺確率を用いた確率的単語分割コーパスの構築とその応用，NLP 若手の会シンポジウム (2006).
4) 海野裕也，坪井祐太：係り受け周辺確率に基づく文節間距離，言語処理学会年次大会 (2010).
5) Graham, N., 中田陽介，森　信介：点推定と能動学習を用いた自動単語分割器の分野適応，言語処理学会年次大会 (2010).

（平成 22 年 5 月 7 日受付）

☆1　文献 5）では被験者実験を行い，文全体よりも単語単位での正解付与の方が作業時間当たりの性能上昇が大きいことを示している．

坪井 祐太（正会員）　yutat@jp.ibm.com
2009 年奈良先端科学技術大学院大学博士後期課程修了．工学博士．現在日本アイ・ビー・エム（株）．

森 信介（正会員）　mori@ar.media.kyoto-u.ac.jp
1998 年京都大学博士後期課程修了．工学博士．現在京都大学准教授．

鹿島 久嗣（正会員）　kashima@mist.i.u-tokyo.ac.jp
2007 年京都大学博士後期課程修了．情報学博士．現在東京大学准教授．

小田 裕樹（正会員）　oda@fw.ipsj.or.jp
1999 年徳島大学博士前期課程修了．工学博士．

松本 裕治（正会員）　matsu@is.naist.jp
1979 年京都大学修士課程修了．工学博士．現在奈良先端科学技術大学院大学教授．