

審査の結果の要旨

氏名 李海博

本論文は「Graph Based Learning Methods For Semantic Relation Extraction (意味的關係抽出のためのグラフに基づく学習手法)」と題し、英文で記されており、8章から成る。

第1章「Introduction (序章)」では、対象とする膨大な Web テキストから意味的關係を有するエンティティ対を抽出する意義、学習法を含む関連手法、本論文の貢献についての概観を述べている。ここで対象とする意味的關係を有するエンティティ対とは、例えば買収・被買収関係にある2企業名、人間に関する事項では人物と出生地といった対である。

第2章「Semantic Relation Extraction (意味的關係抽出)」では、Web テキストからの関係情報抽出の関連手法についてまとめている。大別すると3種の枠組みが存在し、1)関係種別ラベルのアノテーションが付与された大量のコーパスデータから教師付き学習によって識別器を求める伝統的關係抽出の手法、2)一般性のあるパターン或いは関係種別に非依存なヒューリスティックスを用いて、出現する名辞エンティティ間に存在する全ての関係の抽出を図るオープン関係抽出の手法、3)与えられた少数の種(seeds)を基にして大量の Web 等のテキストコーパスから、種と近い関係を持つエンティティ対の具体例を求めるブートストラッピング型関係抽出の手法である。3)の手法は1)と2)の中間的性格を有し、本論文はこのアプローチについての研究を記している。

第3章「Semi-supervised Learning (半教師付き学習)」では、まず上記のブートストラッピング型関係抽出に関係する半教師付き学習法、特に本論文に関係する共トレーニング(co-training)法とラベル伝搬(label propagation)法について記している。この両者は、異なる仮説に基づいている。共トレーニング法は全てのデータ点 $x=(x_1, x_2)$ に対して、ビュー1で訓練された識別器 f_1 が x_1 からラベル1を与え、ビュー2で訓練された識別器 f_2 が x_2 からラベル2を与える場合、このラベル1とラベル2は等しくなるという両立性仮説(compatibility assumption)に基づく。しかし、実用的にはこの仮説が全てのデータ点で十分に成立するとは期待できない。一方でラベル伝搬法は、近隣のデータ点は同一ラベルになるという整合性仮説(consistency assumption)に基づく。本論文で考案している手法は、上記の両立性が成立しない影響を軽減するため、整合性仮説が両立性仮説を以下のように緩めるものとして見なしている。即ち、1つのビューのデータ点は他のビューで連結されたデータ点と同一のラベルを持つ傾向にあるが、必ずしもそうであるとは限らないと見なしている。このアイデアに基づき、第5章で半教師付き学習のマルチビュー学習法を提案している。

第4章は「Framework of Bootstrapping Relation Extraction (ブートストラッピング型関係抽出の枠組み)」であり、拡張とランキングから成る2ステップの考案した手法を記している。抽出対象となる関係(関係を持つエンティティ対とその周辺テキスト文脈パターンの一方又は両方)が与えられた時、拡張機能は対象となる関係を持つエンティティ対と周辺文脈パターンを、ラベルなしの大量のデータセット(Web)から交互に抽出する。ここで抽出されるエンティティ対や周辺文脈パターンの幾つかは、対象とする関係と無関係であったり関係性が弱いことがあるので、抽出されるデータを関係性により選別することが重要になる。このために、ここでは与えられた種との類似性に従って、拡張機能により抽出されたデータをランキングする機能を導入している。

第5章「Graph based Multi-View Learning (グラフに基づくマルチビュー学習)」では、エンティティ対集合と周辺文脈パターン集合をそれぞれのビューとしたとき、整合性仮説を一般化したビュー内(intra-view)整合性とビュー間(inter-view)整合性を用いる手法を提示している。ここで、ビュー内整合性とは、各ビューに於いて、近傍データ点或いは同一構造上のデータ点は同一ラベルを持つ可能性が高いことを意味する。一方、ビュー間整合性では、異なるビューからのデータ点の共起度が高ければ、これらのデータ点は同一ラベルに属する傾向にあると仮定する。共トレーニング流の手法とは異なり、ここでの提案手法は両立性仮説の成立を要しない。異なるビューから得られるデータ点は異なるクラスに属することがあるので、大部分の実問題ではこの両立性仮説は強過ぎる仮説である。共トレーニング法では両立性仮説を用いて、1つのビューから得られた“ハード”なラベルを他のビューから得られるデータへ伝搬させる。これに対しここでの提案手法では、2部グラフ(bipartite graph)上で“ハード”なラベルを伝搬させるのではなく、異なるビューの間でラベルのスコアを拡散させることにより、ビュー間の非両立性の問題を回避できるようにしている。

第6章「Experiment: Improve Bootstrapping Relation Extraction (ブートストラッピング型関係抽出の改善の実験)」では、ブートストラッピング型関係抽出の実験を行い、前章で提示したマルチビュー学習法の性能を評価している。その結果として、関係スコアに基づく半教師付き学習法や、頻度に基づく半教師付き学習法といった既存手法と比較し、提案手法は関係抽出の性能向上をもたらすことを示している。

第7章「Experiment: Semantic Relation Classification (意味的関係識別の実験)」では、提案のマルチビュー学習法をブートストラッピング型の意味的関係抽出に適用している。そして、テキスト中のエンティティ間関係を44種の関係種別によってラベル付けしてテキストが表わす意味を表現する言語であるCDL(Concept Description Language)に基づくコーパスと、評価用の共通的数据セットとして作成されたSemiEval-07を対象にした実験を行い、提案手法は単一ビューによるラベル伝搬法の性能を改善することを示している。また、提案手法はビュー間の相関や特徴の各ビューへの異なる分割に対して頑健であることを示している。

第8章「Conclusions and Future Research (結言と今後の研究)」では、本論文の成果をまとめると共に、今後の研究課題について言及している。

以上を要するに、本論文は種(seeds)となるある関係を有する少数のエンティティ対を与え、Web等の大量のテキスト中からこれらと近い関係を有するエンティティ対の具体例を求めるブートストラッピング型関係抽出の問題に対し、半教師付き学習で従来用いられてきた共トレーニング(co-training)法が依拠する2つのビューでの両立性仮説は必ずしも成立しないという課題を軽減する手法として、2つのビュー領域間で拡張とランキングから成る2ステップの操作を行い、また従来のラベル伝搬法が依拠していた整合性仮説を一般化したビュー内(intra-view)整合性、ビュー間(inter-view)整合性を用い、2つのビュー領域から成る2部グラフ上でハードな関係ラベルを伝搬させるのではなく、異なるビュー間でラベルのスコアを拡散させるマルチビュー学習法を考案、開発している。そして、実験を通して考案した手法による性能改善効果を示している。この研究成果は、少数の関係種別ラベル付きデータに基づき、大量の関係種別ラベルのなしデータを活用する情報処理技術に貢献するものと認められ、情報理工学における創造的実践の観点でも価値が認められる。

よって本論文は博士(情報理工学)の学位請求論文として合格と認められる。