1Q-1

半教師あり学習を用いた述語項構造解析

岸邊 賢太† 横山 晶一†

山形大学大学院理工学研究科†

1.はじめに

自然言語処理の高度化に伴って、意味解析が必要とされる。その１つである述語項構造解析は、文中で事態を表す述語と関係する項を同定することで文章の意味を解析する。現在は、教師あり学習が主に成果をあげている[1,2]。しかしこれらの手法は、トレーニングデータとして大規模なタグ付きコーパスが必要である。このような状況は、述語項構造解析の他分野での利用に障壁となることが考えられる。

そこで、低コストでトレーニングデータを収集するために半教師あり学習[3]の利用が望まれる。

本研究では、TSVM[4]による半教師あり学習を用いた述語項構造解析手法を提案する。

2.研究背景

2.1.述語項構造

述語項構造解析は、文における意味の核である述語とそれに関係する項（日本語においては名詞句）を同定するタスクである。『私が車を運転して帰った。』という文を例に説明する(図 1)。述語は、『運転して（基本形：運転する）』と『帰る』である。それぞれの述語に対する「誰（何）ガ誰（何）ヲ誰（何）ニ」が項となりうる。すると『運転する：私ガ,車ヲ』、『帰る：私ガ』となる。この項と述語の構造を述語項構造という。

図 1:述語項構造例

2.2.関連研究

組み合わせ特徴量と教師あり学習の SVM（Support Vector Machine）を用いた述語項構造解析の研究[1]がある。また、統計的関係学習を用いて文章全体の整合性を取りながら述語項構造解析を行う手法[2]が提案されている。いずれの研究も、NAIST テキストコーパスを最大限に利用し精度向上を図っている。

3.研究内容

3.1.半教師あり学習

大規模タグ付きコーパスは、非常に多くの労力・コストがかかり、構築が容易でない。これは、教師あり学習を利用する上での欠点である。そこで、半教師あり学習を用いる。この手法は、タグ付きコーパスの利用量を削減し、収集コストの低いタグなしデータで補い、トレーニングを行う手法である。意味解析のタスクである語義曖昧性解消において成果をあげている[5]。

3.2.Transductive SVM

TSVM は、半教師あり学習である Transductive Learning を行う SVM である。これは、テスト用のラベルなしデータをトレーニングに使用し、正確なラベル付与を行う学習手法である。SVM についての詳細な説明は省略する。TSVM の主な手順は次のようである。

①少量のタグ付きデータで分離平面を構築
②分離平面からタグなしデータを分類
③境界に十分近いタグなしデータのペアを選択
④ペアラベルの交換　（マージンが大きくなる時）
⑤終了条件が成立すると終了　（②から繰り返し）

3.3.提案手法

本研究では、述語項構造解析の利用に必要なタグ付きコーパス作成コストの低減を図るために半教師あり学習による述語項構造解析手法を提案する。

本手法は TSVM と SVM を用いた 2 段階に分けて分類を行う。手法の手順を以下に示す。

(1)TSVM を用いた分類

ここでは、TSVM を用いてタグなしデータにタグを付与することを行う。

①タグ付与コーパスのトレーニングデータ化
　コーパスから素性値を計算し、トレーニングデータを作成する。
② TSVM によるラベルなしデータの解析
　①とタグなしデータを利用して、タグなしデータの解析を行う。
③解析結果からタグ付加
　タグなしデータにタグを付与する。このデータ

Predicate Argument Structure Analysis using Semi-Supervised Learning

† Kenta Kishibe , Shoichi Yokoyama

† Graduate School of Science and Engineering (Informatics), Yamagata University

Copyright ©2013 Information Processing Society of Japan.
All Rights Reserved.

を以降、タグ付加データと記す。

(2)SVM を用いた分類

ここでは、(1)の結果を利用し増量を図ったトレーニングデータを用いたテストデータの解析を行う。

④ SVM によるテストデータ解析

①及び③によって作成されたデータでトレーニングを行い、テストデータを解析する。

本研究では、述語同定、項同定、格分類（ガ、ヲ、ニ格）を解析する。そのため上記の手順は、個々に行われる。また、形態素および係り受けは、正しく解析されていることを前提としている。

3.4.素性

提案手法では、大量のタグ付きデータを使用しない。そのため、複雑な素性を利用することは、困難である。そのような点を考慮して素性の作成を行う。以下にその一部を示す。

・述語同定

基本情報（基本形、品詞、品詞細分類、活用型、活用形、受身表現、使役表現）

・項同定、格分類

述語・項候補の基本情報、係り受け関係

4.実験

提案手法を評価するため、教師あり学習の SVM を用いた比較手法、提案手法での比較実験を行う。

4.1. 比較手法

比較対象として SVM による教師あり学習手法を実装した。提案手法と同じ素性を利用する。以下本稿では、SVM による教師あり学習手法を比較手法と記す。

4.2.実験設定

半教師あり学習が必要とされる状況を想定した実験を行った。実験方法は以下に示す通りである。

・ ベース記事を設定

・ 以降、記事数を数記事ずつ追加

・ ベース記事をタグあり、追加記事をタグなしとして利用

・ テストデータを別記事として用意し、評価

実験には、NAIST テキストコーパスを利用した。また、記事選択は、全てランダムである。

本稿で示す実験設定は以下のようである。記事とは、複数の文が含まれている文書である。

・ ベースデータ　　：2 記事

・ トレーニングデータ: 4/6/8/10/15/20 記事

・ 実験用データ　　：50 記事

結果の評価には F 値を用いる。

4.2.実験結果

述語同定における実験結果を図 2 に示す。破線がベース記事を用いた比較手法、実線が提案手法である。提案手法は、比較手法とほぼ同等の結果を出している。しかし、トレーニング記事の増加に伴う精度の向上については確認することができない。

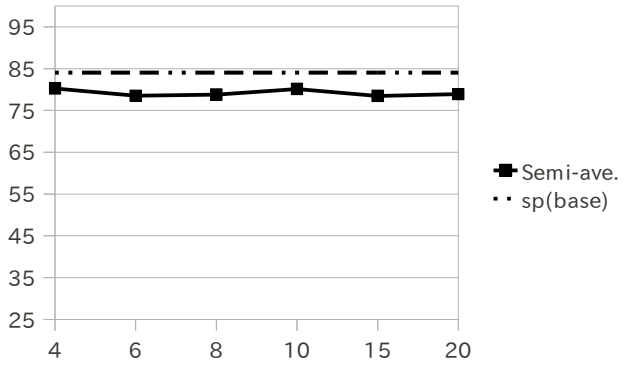

図 2:実験結果 -述語同定-

4.3.考察

今回の実験では、教師あり学習に並ぶ解析を行うことは可能であると考えられる。しかしながら、タグありデータが少量の状況で、タグなしデータの追加によって精度を向上させているかについてはデータ数を増加させて実験を行うことが必要である。

5.おわりに

本研究では、半教師あり学習を用いた述語項構造解析について提案した。手法としては、TSVM を用いた Transductive Learning によってタグなしデータにタグを付与することでタグありデータの増強を図り、解析を行う。結果は、述語同定に関して教師あり学習と同等程度の結果であった。

今後は、格分類における評価、データの選別方法の検討や、アンサンブル学習を用いてより安定した解析が行えるようにすることが必要となると考えられる。

参考文献

[1]平,藤田,永田.”組み合わせ特徴量を用いた述語項構造解析”,言語処理学会第 17 回年次大会発表論文集,2011.

[2]吉川,浅原,松本.”Markov Logic による日本語述語項構造解析”,情報処理学会自然言語処理研究会研究報告,SIGNL-199,2010.

[3]X. Zhu. ”Semi-Supervised Learning Tutorial”, ICML2007, 2007.

[4]T. Joachims. ”Transductive Inference for Text Classification using Support Vector Machines”, Proc. 16th Int. Conf. on Machine Learning, 1999.

[5]井上,斎藤.”ラベルなしデータの二段階分類とアンサンブル学習に基づく半教師あり日本語語義曖昧性解消”,自然言語処理 Vol.18, No.3, 2011.

Copyright ©2013 Information Processing Society of Japan.
All Rights Reserved.