ルールを用いた教師データ自動獲得による競合企業名抽出

谷口元樹　杉原大悟　三浦康秀　大熊智子
富士ゼロックス株式会社　研究技術開発本部
{motoki.taniguchi, daigo.sugihara, yasuhide.miura, ohkuma.tomoko}@fujixerox.co.jp

1. はじめに

営業支援システムの発展に伴い、営業日報などのテキストデータが企業内に大量に蓄積されている。これらの大量のテキストから重要文、商談の成功・失敗事例、因果関係の抽出を行うことで、企業の意思決定に役立てる取り組みが行われている[1,2,3]。

企業間の商談において、顧客は複数の企業に見積もりを要求し、最も価格や条件の良い企業との商談を選択することが多く行われる。したがって、セールスパーソンにとって競合情報は重要であり、顧客から入手した競合情報は営業日報に入力されている。この競合情報を自動抽出することは、他のセールパーソンや販売戦略の決定者にとって有用である。

本論文では、競合情報抽出の一歩として、競合企業名の抽出を行う。競合企業名を固有表現として捉え、営業日報中の競合企業名に対してアノテーションされたコーパスを用意した教師あり学習手法を用いることで抽出することは可能である。しかし、この手法においては、コーパス作成時に三つの課題がある。一つ目の課題は、営業日報中に競合企業名は稀にしか出現せず、大量の教師データを用意するためには人的な負荷がかかることである。二つ目の課題は、営業日報中に出現する競合企業の中には、商談内容によって競合にも非競合にもなる企業があることである。そのため、営業日報中に出現する企業名が自社に対して競合しているかどうかの判断には企業の活動や製品に対する専門的知見が必要となる。三つ目の課題は、営業支援システムを利用しているユーザ企業ごとに、競合企業は異なることである。このため、異なるユーザ企業の営業日報から競合企業名を抽出するためには、コーパスを作りなおさければならないことである。

そこで、本論文では、ルールを用いて教師データを自動獲得することで、アノテーションの人的な負荷や知見といったコーパス作成コストを軽減した競合企業名抽出システムを提案する。

2. 営業支援システムにおける営業日報

営業支援システムは営業プロセスを管理することで、営業活動の効率化を目指す情報システムである。セールスパーソンは営業活動の進捗報告として営業日報をシステムに入力する。営業日報には訪問先、販売商品、商談の進捗状況など様々な項目が存在するが、本論文では活動内容を記入したテキストを対象にする。

営業日報中に企業名が見られる例を表 1 に示す。例文１と２の“××社”のように、いずれの商談においても競合として現れる企業がある。一方で、例文２、３、４の“△△社”のように、ある商談では競合企業として現れ、他の商談ではパートナー企業や販売製品のメーカー企業として現れる企業がある。このように競合であるか非競合であるかの自社との関係性が、商談内容に依存する企業と依存せず確定的な企業がある。自社との関係性が確定的な企業を確定的企業、確定的ではない企業を不確定企業と以後呼ぶ。

表 1　営業日報中に企業名が見られる例文

	例文	競合企業名	非競合企業名
1	××社からは見積もり提示あり。	××社	
2	複合機案件の競合は××社と△△社。	××社 △△社	
3	ペーパーレスについて△△社と同行。		△△社
4	ノート PC の見積依頼あり。 △△社の AB-12 モデル。		△△社

3. ルールを用いた教師データの自動獲得

競合企業名がアノテーションされたコーパスを人手で作成するためには、作成コストが高いことが課題となる。そこで、営業日報に対して企業名抽出を行い、抽出された企業名に対して単純なルールを用いてラベル付けすることで、教師データを自動獲得する。

提案手法では、以下の二つの仮説に基づいている。

1. 競合企業と非競合企業の出現する文脈は異なる。
2. 確定的企業と不確定企業が出現する文脈は類似している。

これらの仮説に基づけば、確定的企業の出現する文

をルールで自動獲得し、教師データに用いることで不確定企業に対して競合か非競合かの判定が可能である。

下記の二つのルールに基づいて、競合であると判断した企業を正例、非競合であると判断した企業を負例とする。

1. 確定的企業名のリストと照合する。
2. 抽出された企業名の語尾に“様”が含まれるものは非競合とする。

ルール１において、確定的企業のうち競合である企業はセールスパーソンにとっては常識として知られているため作成することは容易である。また、確定的企業のうち非競合である企業の例としては自社の関連会社や子会社が挙げられる。ルール２においては、企業名として抽出された固有表現の語尾に“様”が含まれるものは、商談相手となる顧客企業名、もしくはベンダーなどの協力企業名である、との仮説に基づいている。

4. システム概要

図 1 で示すように、提案システムでは２ステップの処理を行うことで、営業日報から競合企業名を抽出する。１ステップ目では、固有表現抽出において一般的な機械学習手法を用いて企業名を抽出する。２ステップ目では、１ステップ目で抽出された企業名を対象に、競合であるかどうかの分類を行う。２ステップ目の競合企業名の分類においては、ルールによる分類と機械学習を用いた競合企業名分類器を併せて用いる。

２ステップ目の機械学習を用いた競合企業名分類器の教師データは、2 章で述べたルールを用いて自動獲得されたデータを用いる。教師あり学習に基づく固有表現抽出手法では、教師データとして競合企業名タグがつけられたフルアノテーションコーパスが必要であった。一方で我々のシステムでは企業タグ付きの簡易アノテーションコーパスでよいことになる。これにより、コーパス作成のコストが大幅に軽減される。これは競合企業名の出現頻度よりも企業名の出現頻度が高いことや企業名かどうかの判断に必要となる知識レベルが低いためである。

4.1 ステップ１　企業名の抽出

１ステップ目の企業名の抽出では、固有表現抽出でよく用いられるConditional Random Field（CRF）[4]を文字単位で利用する。素性は、文字の表層、文字種、形態素の表層、形態素の基本形、形態素の品詞、形態素の活用の種類、形態素の活用形を採用する。また、ルールを用いた教師データの自動獲得時の前処理の企業名抽出においても、同様のものを用いる。教師データは企業名タグがつけられた簡易コーパスを用いる。

4.2 ステップ２　競合企業名分類

教師データの自動獲得において、用いたルールを一段目の抽出結果に対しても適用することで、確定的企業と不確定競合企業に分ける。確定的企業に関しては、ルールに従い、競合企業分類を行う。また、ルールに適合しない不確定企業は次節の競合企業分類器の処理を適用する。また、機会学習を用いた競合企業分類器はSupport Vector Machine[5]を利用する。素性は、同一文中の自立語の基本形のユニグラムとバイグラムを採用する。ただし、対象となっている企業名の字面による影響をなくすために、対象企業名は素性に含めないものとする。

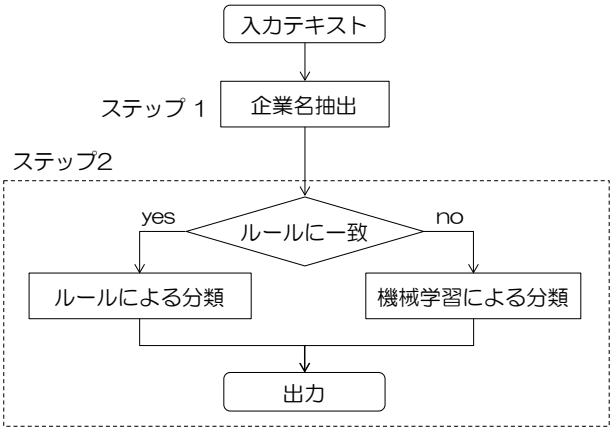

図 1　システムの処理の流れ

5. 評価実験

評価実験にあたっては、弊社の営業支援システムに蓄積された約 130 万件の営業日報のテキストを用いた。

5.1 人手で作成したタグ付きコーパス

3500 件の営業日報をサンプリングし、文中の企業名に対して企業名タグを、さらに競合である場合には競合企業フラグを付けることで、コーパスを作成した。コーパス中に、企業名は延べ数で 2,860 個、競合企業名は延べ数で 536 個存在した。

5.2 ルールを用いて自動獲得された教師データ

競合企業分類器の教師データの自動獲得にあたっては、ルール 1 で必要となる確定的企業名のリストを人手で作成した。確定的競合企業は 5 社、確定的非競合企業は 48 社を選択した。それぞれの企業名の呼称と略称などの異表記含めると、確定的競合企業名 16 個、確定的非競合企業名 136 個をリストに登録した。

5.1 で作成した簡易コーパスで学習した企業名抽出器を用いて、営業日報から企業名を抽出する。その抽出された企業名に対してルールを適用することで教師データを作成した結果を表 2 に示す。ただし、人手で作成したタグ付きコーパスで対象とした 3500 件のテキストは対象テキストから除外した。ル

ール１では、教師データの正例数の方が負例よりも非常に多く、不均衡な状態である。ルール２では負例のみが自動獲得されるために、ルール１＋２ではルール１よりも不均衡な状態が改善されている。また、タグ付きコーパスのタグの延べ数と比較すると、大量の教師データが自動獲得できていることがわかる。

表 2　自動獲得された教師データ

適用ルール	正例数	負例数	合計
ルール１	18,390	3,277	21,667
ルール１＋２	18,390	14,016	32,406

5.3 評価条件

１ステップ目の企業名抽出においては、企業名タグ付きの簡易アノテーションコーパスを学習：評価＝４：１に分割して精度を評価した。

２ステップ目の競合企業分類器の教師データは、2 章で述べたルールを用いて自動獲得されたデータを用いる。自動獲得されたデータは１ステップ目の処理の結果を用いているため、誤りを含んだデータになっている。評価においても同様に、タグ付きコーパスに対して企業名抽出を行ったものを用いた。したがって、仮に競合企業名タグがつけられている語句であっても、１ステップ目で抽出されなかった企業名は２ステップ目の評価データに用いていない。

5.4 比較手法

提案手法の有用性を評価するために、二つの手法と比較を行った。それぞれの手法の特徴を表 3 に示す。

比較手法１は競合企業名を固有表現として捉えた手法である、競合企業名タグ付きのフルアノテーションコーパスで教師あり学習を行ったものを採用した。具体的には、提案手法の１ステップ目の企業名抽出と同様に、CRF を用いた。素性も同様に、文字の表層、文字種、形態素の表層、形態素の基本形、形態素の品詞、形態素の活用の種類、形態素の活用形を採用した。

比較手法２は、自動獲得されたデータを用いて、１ステップの教師あり学習を行ったものを採用した。この手法においても、CRF を用いた。素性も同様に、文字の表層、文字種、形態素の表層、形態素の基本形、形態素の品詞、形態素の活用の種類、形態素の活用形を採用した。自動獲得されたデータの負例は、CRF では全て O タグとして扱われるため教師データとして用いていない。この手法と比較することで、提案手法の２ステップ化と表層情報を素性から除いた効果を測ることができる。

表 3　各手法の特徴

手法	学習コーパス	学習手法
比較手法 1	フルアノテーション	CRF
比較手法２	簡易アノテーション、自動獲得データ	CRF
提案手法	簡易アノテーション、自動獲得データ	CRF＋SVM

5.5 評価結果

表 4、表 5 に精度の評価結果を示す。表 5 の精度は２ステップ目の機械学習による分類処理単独での精度を表している。機械学習による競合企業分類では、ルール１にルール２を追加することで F 値が 10pt 向上している。これは表 2 で示されているように、ルール２を追加することで教師データの正例と負例の数の不均衡が緩和されることで、精度が改善されていると考察される。

表 6 にシステム全体の精度を示す。提案手法の２ステップ目の教師データの自動獲得においては、ルール１＋２を採用した。提案手法は比較手法１と比較して、Precision と F 値は低いものの、Recall はほぼ同等レベルに達している。一方で、比較手法 2 と比較すると提案手法は、Precision は低いものの、Recall と F 値では上回っている。これは比較手法２では、自動獲得された確定的企業の表層情報を学習しているため、確定的競合企業かどうかの分類器として動作しており、不確定企業は抽出できていないためであると考えられる。

精度改善のため提案手法における出力のエラー分析を行った。その結果、“××社、△△社、□□社。”のように文長が短く、分類の手がかりとなりえる情報が少ない文で誤りが多い傾向であった。２ステップ目の SVM による競合企業分類においては、素性は文中のユニグラムとバイグラムという非常に単純なものであった。そのため、文長が短い場合には、誤りが多いと考えられる。この問題への対策としては、一文単位ではなく、複数文や文書単位での処理が考えられる。ある企業名が競合かどうかは、多くの場合は一つの営業日報中で変化することはないと考えられるので、前後の文や文書単位での大域的な素性を考慮することで、精度を改善することができると考えられる。また、企業名は複数の企業名が並列して書かれる事が多い。並列された企業が競合であるかどうかは一致すると考えられるので、並列関係にある企業を一括して分類することで分類の手がかり情報を増やすことできる。

表 4　1ステップ目の企業名抽出の精度

	Precision	Recall	F 値
企業名抽出	0.81	0.66	0.73

表 5　2ステップ目の競合分類器の精度

適用ルール	Precision	Recall	F 値
ルール１	0.28	0.95	0.43
ルール１＋２	0.41	0.77	0.53

表 6　システム全体の精度

手法	Precision	Recall	F 値
比較手法１	0.70	0.52	0.60
比較手法２	0.76	0.30	0.43
提案手法	0.43	0.55	0.48

6. おわりに

本論文では、人手によるタグ付きコーパス作成のコストを軽減した上で、大量の営業日報テキストから競合企業名を抽出する手法を提案した。提案手法では企業名を抽出する処理と抽出された企業名が競合企業であるかどうかを分類する処理の２段階処理で構成される。２ステップ目の処理の競合企業の分類器の学習においては、大量の教師データを単純なルールによって自動獲得できることを示した。評価実験においては、Precision はベースライン手法に及ばないものの、Recall は同等レベルであることを示した。

今回の実験では、弊社の営業日報を対象に競合企業名抽出を行った。他社の営業日報から同様の抽出を行うことを考えた場合、比較手法１はその会社に合わせたフルアノテーションコーパスの作成が再度必要となる。しかし、提案手法であれば、企業名タグ付きのコーパスを作成し、ルール１の確定的企業リストを変更することにより、容易に競合企業名を抽出できる点は有用である。

参考文献

1 杉原大悟, 大熊智子, 佐竹功次, 三浦康秀, 服部圭悟. 営業支援システム内に蓄積されたテキストデータからの課題記述文抽出.情報通信学会技術研究報告 言語理解とコミュニケーション 112(196), pp.7-12, 2012.

2 柴田親男, 松田純一, 小泉敦子, 森本康嗣. 企業における非定形文書の活用促進事例: 営業日報へのテキスト分析技術の適用(自然言語処理技術による情報マネジメントの実際)(<特集> 自然言語処理の高度化による知的生産性の向上). 情報処理, Vol.44,No.10, pp.1022-1027, 2003.

3 市村由美, 鈴木優, 酢山明弘, 折原良平, 中山康子. 日報分析システムと分析用知識記述支援ツールの開発. 電子情報通信学会論文誌. D-II, 情報・システム, II-パターン処理 J86-D-II(2),pp. 310-323, 2003.

4 J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceeding of the 18th International Conference on Machine Learning, 2001.

5 V. Vapnik, "The Nature of Statistical Learning Theory", Springer, (1995).