事態間知識における項の対応付け学習

小浜 翔太郎 † 柴田 知秀 †‡ 黒橋 禎夫 †‡
† 京都大学大学院情報学研究科 ‡ 独立行政法人 科学技術振興機構 CREST
{kohama, shibata, kuro}@nlp.ist.i.kyoto-u.ac.jp

1 はじめに

機械学習の進歩や大規模テキストの利用により，形態素解析や構文解析は高精度で行うことができるようになってきた．しかし，**省略・照応解析**を高精度で行うことはまだ難しい．省略・照応解析に必要となる知識のひとつに述語と項の関係があり，これは**格フレーム**という形で大規模に自動獲得されている [1]．しかしながら，格フレームだけでは照応先を解析するには難しい例も存在する．例えば「グーグルはモトローラを買収した。彼ら が破綻していたからだ。」における「彼ら」が「グーグル」と「モトローラ」のどちらを参照しているかを解析する場合である．上記の「彼ら」は述語と項の関係のみで解析するのは難しいが,「ある会社 A が破綻すると，別の会社 B が A を買収することが多い」という知識があれば「彼ら」が「モトローラ」を参照していることが解析できる．したがって，

「A が破綻する $\Rightarrow$ B が A を買収する」

のような２つの良く起こる出来事とそれらの間にどのような項の対応付けがあるかという知識（以降，**事態間知識**と呼ぶ）が省略・照応解析の重要な手がかりとなる．日本語の事態間知識を大規模に獲得する手法として，Shibata ら [2] はアソシエーション分析を用いたよく共起する事態対の抽出と，格フレームを用いた項の対応付けを２段階に分けて行う手法を提案している．

本研究では，Shibata らの手法をもとによく共起する事態対の抽出を行い，ヒューリスティックなルールによって行われていた項の対応付けを機械学習によって行う．また，日本語の事態間知識に対する評価セットの構築法も提案する．Shibata らは 100 個の知識をランダムサンプリングして自ら評価を行っている．この評価手法では高品質な評価ができる一方で，大規模に知識を評価することは難しい．そこでクラウドソーシングによって，抽出された事態対の共起に対する評価と項の対応付けの正解データ作成を行った．アノテーションの信頼性という，クラウドソーシングによる言語リソース作成における一般の問題に対しては，２段階でタスクを実施することと，Whitehill らが提唱した EM アルゴリズムに基づくラベルの確率値推定手法 [3] で対処した．

以降，２節で日本語における事態間知識獲得の既存手法，３節で学習による項の対応付け，４節でクラウドソーシングを用いた対応付けデータの作成，５節では提案手法の評価と考察について述べる．

2 日本語における事態間知識獲得の既存手法

英語では照応解析が高精度で解析可能となってきており，照応解析は知識獲得や様々なタスクに重要な基礎処理として利用されている．英語で提案されている事態間知識を獲得する手法は，共参照を手掛かりとする手法が多い [4]．しかしながら，英語と比較すると日本語では照応詞の性別や単複の情報がないうえに，照応詞の省略がよく起こるために，共参照を高精度で解析することは現在も難しい．そのため，英語での手法をそのまま日本語の事態間知識獲得に利用することは難しい．

そこで Shibata ら [2] は日本語において事態間知識を獲得するために，１段階目にアソシエーション分析を用いてよく共起する事態対の抽出を行い，２段階目に格フレームを用いて事態間の項の対応付けを行う手法を提案した．Shibata らの手法では**述語項構造**を事態の単位として扱っている．項は格とその格で述語と関係をもつ語から構成され，述語項構造は述語と，述語と係り受け関係をもつ項で構成される構造とする．例えば「太郎が花子にお金を貸す」という文に現れる述語項構造は,「貸す (ガ格:太郎, ニ格:花子, ヲ格:お金)」と表すことができる．この手法によって Shibata らはおよそ 100,000 個のユニークな事態と 340,000 個の事態間知識を得ている [5]．

１段階目のアソシエーション分析は Apriori[6] というアルゴリズムで行われる．Apriori はアイテムセット集合の中から頻出アイテム集合とアソシエーション

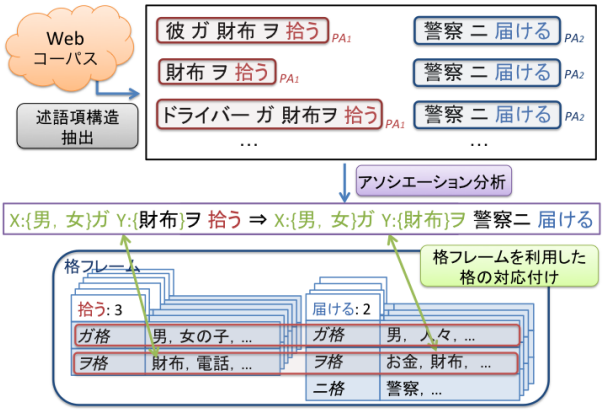

図 1: Shibata らによる事態間知識作成の流れ

ルールを効率よく獲得するアルゴリズムの 1 つである．Shibata らは，項と述語をアイテムとみなし，コーパスで係り受け関係を持つ述語項構造対をアイテムセットとみなすことによって，よく共起する事態対をアソシエーションルールとして効率的に獲得する．この段階で得られた事態対には，コーパスでよく省略される項は残っていない．例えば，図 1 における 1 段階目で獲得した事態対は「財布を拾う ⇒ 警察に届ける」であり，よく省略される主語や目的語の項は残っていない．

2 段階目では，格フレームを用いて事態間の項の対応付けを行う．格フレームは，述語がテキスト上でどのような語と一緒に現れるかを格ごとに整理した言語リソースである．述語の多義性を扱うために格フレームは語義ごとにクラスタリングされており，ある述語の格フレームは番号付きで複数存在する．格フレームを用いて項の分布を推定することで格フレーム間で項の類似度を計算し，1 段階目では残っていない項に対しても対応付けをもった事態間知識を獲得する．例えば，図 1 において最終的に得られる事態間知識は「X が Y を拾う ⇒ X が Y を警察に届ける」であり，1 段階目の段階では残っていなかった「届ける」の「が」と「を」が対応付いている．

3 学習による項の対応付け

提案手法の説明に必要な用語を導入する．ある事態間知識 $ek : PA_1 \Rightarrow PA_2$ に対して，事態 PA_1, PA_2 の述語を $pred_1, pred_2$ とする．本研究では事態間知識のもととなる事態対は Apriori を用いて獲得する．ある事態間知識に対して，その事態対が獲得されるもととなる述語項構造対の集合を，その事態間知識の**支持述語項構造対**と呼ぶ．例えば，図 1 における「X が Y を拾う ⇒ X が Y を警察に届ける」という事態間知識の支持述語項構造対には，「拾う (ガ格:彼, ヲ格:財布), 届ける (ニ格:警察)」が含まれる．

次にクラウドソーシングによって作成した対応付けの正解データを使って，対応付けを学習するモデルについて述べる．学習は**最大エントロピー法**で行う．ある事態間知識 $ek : PA_1 \Rightarrow PA_2$ が与えられたときに，cf_1, cf_2 を $pred_1, pred_2$ に対応する格フレームとし，a を項の対応付け，ps を ek の支持述語項構造対としたとき，以下のように事後確率をモデル化する．

$$P(cf_1, cf_2, a|ek, ps; \Lambda) = \frac{\exp\{\Lambda \cdot \mathbf{F}(cf_1, cf_2, a, ek, ps)\}}{Z(ek, ps)}$$

$$Z(ek, ps) = \sum_{\{cf_1, cf_2, a\} \in C(ek, ps)} \exp\{\Lambda \cdot \mathbf{F}(cf_1, cf_2, a, ek, ps)\}$$

ここで Λ は学習した重み，$\mathbf{F}(cf_1, cf_2, a, ek, ps)$ を素性ベクトル，$C(ek, ps)$ は ek, ps が与えられたときの格フレーム番号と対応付けの候補の集合とする．このモデルを用いて対応付けの予測を行うには，事後確率が最大になるような対応付けを求める．

$$(\hat{cf_1}, \hat{cf_2}, \hat{a}) = \underset{cf_1, cf_2, a}{\mathbf{argmax}}\ P(cf_1, cf_2, a|ek, ps)$$

次にこのモデルの重みの推定法について説明する．素性に用いるすべての情報がデータに付与されている場合の学習とは異なり，今回の学習データには格フレーム番号の情報は付与されていない．これは省略解析における，学習データに省略されている項は付与されているが，格フレームの番号は付与されていない状況に類似している．そこで Sasano ら [7] が提案した，格フレーム番号の推定とパラメータの推定を交互に行う省略解析のモデルを参考に，N 個の訓練データ $\{(a^{(1)}, ek^{(1)}, ps^{(1)}), ..., (a^{(N)}, ek^{(N)}, ps^{(N)})\}$ から重み Λ を次の手続きにしたがって学習する．

1. Λ を [0.0, 1.0] でランダムに初期化する
2. 訓練データすべてに対して，現在の Λ を使って格フレーム番号を推定する．

$$(\hat{cf_1^{(n)}}, \hat{cf_2^{(n)}}) = \underset{cf_1^{(n)}, cf_2^{(n)}}{\mathbf{argmax}}\ P(cf_1^{(n)}, cf_2^{(n)}, a^{(n)}|ek^{(n)}, ps^{(n)}; \Lambda)$$

この際に 2 つの格フレームのすべての組み合わせを考慮すると非常に数が大きくなってしまうので，支持述語項構造対と格フレームにおける項の類似度を計算し候補を減らす．推定された格フレーム番号が前回と比較してすべて同じ場合，または更新の上限回に達した場合に学習を終了する．そうでない場合は step 3. に進む．

3. step 2. で推定した格フレーム番号と訓練データを用いて L2 正則化項付き対数尤度 L_A を最大化する Λ を求めて step 2. へ戻る．その際，対応付けが $a^{(n)}$ でないものをランダムに選んで訓練データに負例として追加する．

$$L_A = \sum_{n=1}^{N} \log P(\hat{cf_1^{(n)}}, \hat{cf_2^{(n)}}, a^{(n)} | ek^{(n)}, ps^{(n)}; \Lambda) - \alpha \|\Lambda\|^2$$

step 2., step 3. ともに尤度は単調に増加するので収束は保証されるが，最大値が求まる保証はないため，複数の初期値を用いて推定を行い，それらの平均を最終的な推定値として使用する．実装には Classias[8] を利用した．

素性としては，3 種類の素性を利用した．対応付けされている項ペアに対しては，格フレーム間と支持述語項構造対それぞれでの項の類似度を対応付きやすさを示す実数素性として，対応付けされていない項ペアに対しては，対応付けされていないことを示すバイナリ素性を利用した．

4 クラウドソーシングによる対応付けデータの作成

本研究ではクラウドソーシングによって抽出した事態対の共起に対する評価と項の対応付けの正解データ作成を行った．クラウドソーシングとは，オンラインの不特定多数の人に仕事を頼んで成果物を得るプラットフォームやプロセスのことをいう．クラウドソーシングを利用することで安価で素早く評価やアノテーションを行うことができる．一方で，クラウドソーシングによる言語リソースの作成はアノテーションの信頼性が低いという問題点がある．そこで本研究では 2 段階でタスクを実施することと，Whitehill らが提唱した EM アルゴリズムに基づくラベルの確率値推定手法 [3] で対処する．タスクを 2 段階で行う事で，1 段階目のタスクの誤ったアノテーションを 2 段階目タスクで訂正することで品質を保証する．Whitehill らのモデルでは，設問の真の答えを潜在変数とし，ワーカの能力と問題の難易度をパラメーターとした EM アルゴリズムで推定する．多数派に投票することの少ないワーカの投票の重みが低くなり，単純な多数決に比べアノテーションの品質を保証することができる．

Apriori で得られた事態対 $PA_1 \Rightarrow PA_2$ に対して「が」「を」「に」「で」格で関係をもつ項を対象に，対応付けのアノテーションを行うタスクについて順に説明する．まず 1 段階目のタスク[1]では，ワーカに Apriori で得られた事態対がよく共起するかを「はい」「いいえ」という選択肢のどちらかを選んでもらった．そして PA_1 のそれぞれの項が PA_2 の項に対応するかを選んでもらった．例えば PA_1 の「が」については「A が $pred_1$ と言えたとすると A と $pred_2$ を使った正しい文を教えて下さい」という問題文で「A が $pred_2$」「A を $pred_2$」「A に $pred_2$」「A で $pred_2$」「どれも適切でない」という選択肢のいずれかを選んでもらった．PA_1 の「を」「に」「で」についても同様に問題文を作成し選択肢を選んでもらった．

2 段階目のタスク[2]は，1 段階目のタスクの結果で得られた結果に対して EM アルゴリズムで推定した確率値が高い対応付けだけを対象に行う．1 段階目の結果，事態間知識 $PA_1 \Rightarrow PA_2$ に対して PA_1 の $case_1$ と PA_2 の $case_2$ が高確率で対応付くと推定された場合，「A $case_1$ $pred_1$ $\Rightarrow$ A $case_2$ $pred_2$ という 2 つの出来事がどれくらい連続して起こるかを選んで下さい．ただし A には共通の名詞が入るとし，どちらかの出来事が日本語として不自然な場合は 起こらない を選んで下さい」という問題文で「よく起こる」「たまに起こる」「起こらない」という選択肢のいずれかを選んでもらった．

これらのタスクをランダムサンプルした 1,500 個の事態間知識について Yahoo!クラウドソーシングを利用し行った．すべてのアノテーションは 1 つあたり 10 人につけてもらった．[3]

5 評価と考察

本研究ではクラウドソーシングによって作成した対応付けの正解データを用いた事態間知識の新たな評価法を提案する．機械翻訳の分野でアライメントの評価に用いられる **Recall, Precision, AER (Alignment Error Rate)** という評価指標を事態間知識の評価に用いる．AER では正解データを更にアノテータの評価に合わせて Possible と Sure に分ける．A をシステム出力，S を Sure，P を Possible (ただし $S \subseteq P$) とすると，Recall, Precision, AER は以下の式で表される．

$$\textbf{Recall} = \frac{|A \cap S|}{|S|}, \ \textbf{Precision} = \frac{|A \cap P|}{|A|},$$

$$\textbf{AER} = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}$$

[1] http://crowdsourcing.yahoo.co.jp/request/detail/3588382294

[2] http://crowdsourcing.yahoo.co.jp/request/detail/3588398818

[3] アノテーションにかかった時間はそれぞれ 1.5 時間，7 時間で，費用は約 50,000 円であった．

表 1: 得られた事態間知識

Apriori で得られた事態対	Gold data		ベースライン	提案手法
	Sure	Possible		
熱戦を繰り広げる ⇒ 優勝する	**が** → **が**, **で** → **で**	を → で	に → が, **で** → **で**	**が** → **が**, **で** → **で**
頁を開く ⇒ 表示される	**を** → **が**, **で** → **で**		を → に, に → が	**を** → **が**, **で** → **で**
切手を貼る ⇒ ポストに入れる	**が** → **が**, に → を		が → で, に → に	**が** → **が**, で → で

Recall, Precision は大きいほど，AER は小さいほど，性能が良いことを示す．この指標を使うことによって，自動的に対応付けの評価を行う事ができる．

第 1 段階目のタスクにおける事態対の共起に関する問に対して１０人中６人以上が「はい」と答えており，かつ少なくとも 1 つは正解の対応付けが付いた 937 データを評価に用いた．Sure, Possible ラベルは２段階目のタスクの EM アルゴリズムによる確率に従って付与した.「よく起こる」,「たまに起こる」の確率が最も高い場合に，それぞれ Sure, Possible のラベルを付与し「起こらない」の確率が最も高い場合にはラベルを付与しない．ベースラインは，Shibata らが用いた格フレームにおける項の類似度による対応付けとする．提案手法の対応付けの学習は，更新の上限を 15 回としモデルは 3 回の推定結果の平均とした．学習による対応付けの結果は 5 分割交差検定の結果である．結果， AER がベースラインに比べて提案手法では 34 ポイント改善した．実験で得られた対応付けを表 1 に，実験結果を表 2 にまとめた．

格フレームにおける項の類似度による対応付けに比べ学習による対応付けは学習データ全体の傾向を反映する．そのため実際は対応づいているがコーパスではよく省略されるために格フレームにおける項の類似度が低くなりがちである「が → が」や「で → で」が正しく対応付けできる傾向がある．しかし同じ格が対応付く傾向が強いため，事態間で格が交替して対応付いている場合に誤った対応付けをすることが多い．

表 2: 実験結果

Sytem	Recall	Precision	AER
ベースライン	0.26	0.22	0.76
提案手法	0.59	0.57	0.42

6 まとめと今後の課題

日本語の事態間知識における項の対応付けを機械学習によって行った．またクラウドソーシングによって作成した正解データによって手法の有効性を確認した．今後の課題としては，有効な素性の抽出や学習アルゴリズムの変更によって格交替が起きている場合に対処すること，照応解析のテストデータでの有効性を確認することが挙げられる．

謝辞

本研究は科学技術振興機構 CREST「知識に基づく構造的言語処理の確立と知識インフラの構築」の支援のもとで行われた．

参考文献

[1] Kawahara, D. and Kurohashi, S.: A Fully-lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis, *In Proc of. NAACL HLT* (2006).

[2] Shibata, T. and Kurohashi, S.: Acquiring Strongly-related Events using Predicate-argument Co-occurring Statistics and Case Frames, *In Proc. of IJCNLP* (2011).

[3] Whitehill, J., fan Wu, T., Bergsma, J., Movellan, J. R. and Ruvolo, P. L.: Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise, *In Proc. of NIPS* (2009).

[4] Chambers, N. and Jurafsky, D.: Unsupervised Learning of Narrative Event Chains., *In Proc. of ACL* (2008).

[5] Shibata, T., Kohama, S. and Kurohashi, S.: A Large Scale Database of Strongly-related Events in Japanese, *In Proc. of LREC* (2014).

[6] Agrawal, R., Imieliński, T. and Swami, A.: Mining Association Rules Between Sets of Items in Large Databases, *In Proc. of SIGMOD* (1993).

[7] Sasano, R. and Kurohashi, S.: A Discriminative Approach to Japanese Zero Anaphora Resolution with Large-scale Lexicalized Case Frames, *In Proc. of IJCNLP* (2011).

[8] Okazaki, N.: Classias: A Collection of Machine-Learning Algorithms for Classification (2009).