E-006

類推に基づいた類似分野における知識生成システム
Knowledge generator for similar domain based on analogy

伊田　政樹†　　坂元　盛浩†　　中嶋　宏†
Masaki IDA　Morihiro SAKAMOTO　Hiroshi NAKAJIMA

1. はじめに

因果関係に関する知識は，質問応答システムや原因分析システムなどのエキスパートシステムにおいて重要な知識のひとつである．近年では，電子的な大規模コーパスから自動的に因果関係を抽出する手法が提案されている[1,2,3,4]．エキスパートシステムの応用例として，過去の事例に基づいた原因分析システムを想定した場合，知識が存在しないために原因分析ができなくなる問題がある．この問題は未経験の事例に関する知識を生成できないため生じている．重大な事故や故障は発生頻度が低いため，知識を生成できないことが多い．これに対し，人間(熟練者)は，これまで経験したことがない問題に対しても，過去の類似した経験を基に問題解決を行っている．この手法を知識生成システムとして実現することで，エキスパートシステムの応用範囲拡大が期待できる．

本稿では，類推に基づいて類似した事例から因果知識を生成する手法について提案し，試作した知識生成システムについて述べる．

2. 類推に基づいた知識生成

類推(Analogy)とは，未知の状況の問題解決において既知の類似した状況を利用する認知活動をいう[5]．コーパスから類似分野の知識を生成する手順は，(1) 事象を抽出する，(2) 因果関係を抽出する，(3) 当該分野の因果関係知識を類推する，という 3 ステップで実現される．以下，それぞれについて述べる．

2.1. 事象の抽出

因果関係の「因」または「果」となる現象の単位を事象と定義する．事象は文書中に記述されている現象を因果関係の連鎖としてとらえたときの単位となる．ここでは，格フレーム解析を行い，述語(動詞，形容詞，形容動詞，サ変名詞)に着目し，述語と述語に係る用語の組として事象を抽出する．

2.2. 因果関係の抽出

文書中の因果関係は，接続詞等によって明示的に示される場合と示されない場合がある．ここではその両者に対応するため，以下の 2 手法により事象相互間の因果関係を抽出する．

(1) 識別子に基づいた因果関係抽出

接続詞等により明示的に因果関係が示されている場合，その前後の事象間に因果関係が存在すると推測できる[1]．たとえば，「家屋の倒壊によって死亡した．」という事例においては，識別子「によって」を鍵に事象「家屋の倒壊」と事象「死亡した」の間に因果関係が存在することを推測できる．

(2) モデル学習に基づいた因果強度推定

事象間の因果関係の有無を確率モデルとして学習し，任意の事象間の因果関係の強さを確率モデルに基づいた尤度として因果関係を構築する．[4]や[6]においても任意の事象間の因果関係の強さを確率モデルとして取り扱う試みがなされているが，各々の事象を離散値として扱っているために学習データに出現しない事象の因果関係を抽出できない問題点がある．

確率モデルとして扱うにあたり，離散値である事象を連続量として空間上にマッピングする．単語集合を複数(N 個)定義し，それぞれの単語集合と事象との類似度として N 次元空間内にマッピングする．また，単語集合は辞書分類に基づいて定義する．類似度は単語集合および事象(述語)のワードベクトル間のコサイン距離で求める．ワードベクトルとは，対象となる語の文脈語の集合を頻度情報として数値化し，対象となる語の意味表現を文脈語とその頻度で表現したベクトルである．文脈語の頻度は，事象(述語)の係受けを調べ，係る単語を文脈語としてコーパスから頻度情報を得る．マッピングされた事象に対して，因果関係の有無を統計的に学習する．

任意の事象の組を入力として，因果関係の強さを確率モデルから尤度として算出することができる．

2.3. 因果関係の類推

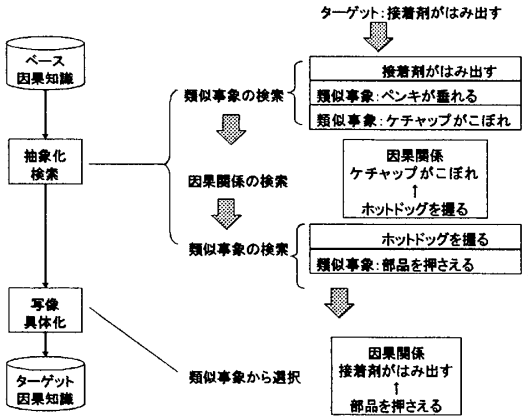

図 1 因果関係の類推

類推は抽象化・検索・写像・具体化の 4 ステップで実行される[5]．ここでは図 1 に示す通り，抽象化・検索ステップと写像・具体化ステップに分けて考える．前者は因果関係の意味や構造の類似性に基づいて検索対象に関連した因果関係を探索する問題であり，後者は探索された因果関係を対象分野の事象に置換する問題である．いずれも事象のワードベクトル間のコサイン距離として類似度を求めることができる[7]．前者では，事象(述語)に係る単語を文脈語とした頻度ワードベクトルをコーパスから求め，ワードベクトル間のコサイン距離を求めることで事象の意味や構造の類似性を算出できる．類似した事象が関係する因果をベースの因果関係から検索し連結する

†オムロン株式会社 センシング＆コントロール研究所，
OMRON Co. Sensing & Control Technology Laboratory

ことで，類似した因果関係に基づいた問題解決を行う．一方後者は，文書中に共起する用語すべてを文脈語としてワードベクトルを生成することで，事象の属する分野の類似性を求めることができる．分野の類似性に基づいて類似した事象から最もターゲットの分野に近い事象を選択することでターゲットに近い因果関係を生成する．

3. 評価実験

3.1. システム構成

実験に用いたシステム構成を図 2 に示す．

前処理として，係り受け解析器として南瓜[8]を用いる．

因果関係抽出における識別子として「ため」「により」を用いる．識別子は失敗知識データベース[9]を用いた予備実験により選定した．事象を空間上にマッピングする際の単語集合は，日本語語彙大系[10]の名詞，固有名詞意味属性大系の分類に従い，バランスされるように 12 分類となるよう予備実験で求めた．因果関係の統計モデルは混合正規分布でモデル化し，EM アルゴリズム[11]により学習を行った．混合数 12，対角共分散のみ扱う．

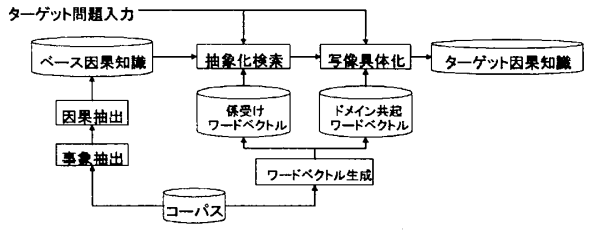

図 2 システム構成図

3.2. 学習データおよび評価データ

学習データおよび評価データとして，毎日新聞データ(CD-毎日新聞 2003, 2004, 2005 年版)[12]および失敗知識データベースを用いる．評価データとして失敗知識データベースの電気電子情報分野，食品分野，自動車分野の 3 分野のデータを用い，それ以外の失敗知識データベースと毎日新聞データを学習データとして用いる．

学習データは，(1) 失敗知識データベース(評価用を除く)1110 事例，(2) (1)に毎日新聞データ 15000 記事を追加，(3) (1)に毎日新聞データ 30000 記事を追加，の 3 種について実験を行い，学習データ量の差による性能を比較する．

評価データは失敗知識データベースの 3 分野を用いる．各分野計 29 件の原因事象に相当する部分を入力文として与えて結果の予測を行う．また，別の計 29 件について結果事象に相当する部分を入力文として与えて原因の推定を行う．

3.3. 実験結果

結果予測，原因推定あわせて 58 件の入力に対し，14 件で意味のある類推結果を得た．うち 3 件で有益な因果関係を類推できた．類推に成功した例を表 1 に示す．

学習データ量と類推性能の関係について表 2 に示す．学習データが少ない場合には適切な因果関係がベースの因果知識になかったために類推失敗となっているのに対し，学習データが多くなるにつれ不適切な因果関係をベースの因果関係から検索したために類推失敗となる傾向が見られた．また，写像・具体化の際に妥当な分野への絞り込みにおいて失敗している例が見られた．

表 1 類推に成功した例

	分野	
入力文	電気	分析用水素元弁を閉め忘れた
結果予測	化学	→出口弁を絞る→タンク内圧力がタンク強度を上回る→過大になる
入力文	食品	アルコール類の容器が落下した
結果予測	石油	→デッキスキンを破壊する→流出する
入力文	自動車	戻り配管のクランプが破損した
原因分析	機械	テーパピンへ変更される→対応が十分に検討されない→誤操作を招く

表 2 学習データ量と類推性能

学習データ	失敗 DB	+新聞 15000	+新聞 30000
類推可能	12	14	13
うち妥当な結果	3	3	3

4. まとめ

本稿では，類推に基づいた知識生成を行うことで類似分野のコーパスから知識を自動的に生成する事が可能であることを示し，システムの試作を行った．評価実験により，類推に基づいて有益な因果関係を生成することを確認した．今後の課題として，類推精度を向上と大規模データを用いた評価の実施が挙げられる．

参考文献

[1] 乾ほか，接続詞「ため」に基づく文書集合からの因果関係知識の自動獲得，情処論誌 Vol.45, No.3, pp.919－933 (2004).

[2] Girju et al., Mining Answers for Causation Questions, Proc. the AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases (2002).

[3] 鳥澤，「常識的」推論規則のコーパスからの自動抽出，言語処理学会発表論文集，pp.318－321(2003)

[4] 山田ほか，クローズドキャプションを対象とした因果関係知識抽出の検討，FIT2005，pp.113－114

[5] 松原ほか，アナロジー入門，情報処理学会誌，Vol.34, No.5, pp.522－535(1993).

[6] 乾ほか，因果関係知識獲得のための隠れ変数モデル，言語処理学会第 12 回年次大会発表論文集(2006)

[7] Takagi et al., A Trial for Data Retrieval Using Conceptual Fuzzy Sets, IEEE Trans. Fuzzy Systems, Vol.9, No.4, pp. 497－505(2001).

[8] 工藤ほか，チャンキングの段階的適用による係り受け解析，情処論誌，Vol.43, No.6, pp.1834－1842(2002)

[9] 失敗知識データベース，(独)科学技術振興機構，http:// shippai.jst.go.jp/

[10] 日本語語彙大系 CD－ROM 版，岩波書店(1999)

[11] Dempster et al., Maximum likelihood from incomplete data via the EM algorithm, J. of Royal Statistical Society Series B, Vol. 39, pp. 1-38(1977).

[12] CD-毎日新聞データ，日外アソシエーツ(株)