D-015

スキーマ構成文字列とワードネットの自動対応付け辞書を用いた語義ベース・スキーマ照合

Matching Abbreviated Multi-Lingual Schemata : Assignment of Schema Element to WordNet

佐藤 彰洋† 谷垣 宏一† 柴 光輝† 山足 光義†
Akihiro Sato Koichi Tanigaki Mitsuteru Shiba Mitsuyoshi Yamatari

1．はじめに

システム統合，データベース統合においては，既存データベースと新データベースの対応関係を把握するために人手で設計作業を行う必要があり，多大なコストを要する．これに対しスキーマ照合により，項目名（データベースのカラム名）に対し，編集距離，部分一致，接尾，接頭，親子関係といった様々な情報による判別を行い，それらを総合判断してデータベース対応関係の自動推薦を行う手法が存在する[1]．しかし，項目名は文字列長が制限されているため任意の短縮表記が使われ，日英語の混用による曖昧性の増加と相まって，スキーマからの情報だけでは同一性が判断できないものも少なくない．

これに対し，汎用シソーラスを類義語辞書として利用し，スキーマからの情報だけでは照合困難なスキーマの自動推薦精度を向上させるという手法が存在する．ただし，WordNet に代表される汎用シソーラスは概念辞書であり，スキーマ中で意図された語義以外の語義も含まれている（語義の曖昧性）．そのため，本稿では，語義曖昧性解消（Word Sense Disambiguation : WSD[2]）により対象スキーマに適した類義語を含む形にした類義語辞書を用いてスキーマ照合を行う手法を提案する．

また，実応用を考えるとき，スキーマ照合では顧客の情報システムからスキーマ情報を取得する必要があり入手可能なデータ量が制限されること，出現する語義が業種や設計者に依存しデータの流用が困難であることから，WSD に十分な量のデータを利用することはコストとのトレードオフになる．従って，WSD の精度向上がスキーマ照合の精度向上に繋がるかを把握することは重要な課題である．本稿では上記の課題を解決するため，WSD の精度とスキーマ照合精度の関係を評価する手法を提案する．さらに既存のスキーマデータを用いた評価実験の結果と考察について述べる．

2．語義ベース・スキーマ照合方式

本稿で述べる手法は，スキーマ照合対象となるスキーマに関して WSD 済みの類義語辞書を作成するフェイズと，スキーマ照合結果を定量化するフェイズからなる．以降，各フェイズについて説明する．

類義語辞書作成

WordNet の語義定義の完全一致は客観性のある類義語定義と考えられるが，応用上の観点からは細かすぎる面がある．WSD 評価型ワークショップ Senseval/SemEval では，語義をより粗い単位(coarse-grained)に纏め直したマッピングが提示され，評価に使われている[3]．そこで本評価では，語義間のリンク数による距離が $d(=1)$ 以内であれば類義語であるとして，粗い粒度の辞書を機械的に生成する．

本評価では，評価対象となるスキーマに関する正しい対応付けをもとに，WordNet から正しい語と語義の対応付けを取得し，ランダムにノイズを入れて誤りを含む WSD を模擬する．特定の誤り傾向は仮定せず各語・各語義において独立・一様に誤りが発生するとし，再現率と適合率がほぼ等しくなるようノイズを入れて類義語辞書を作成する．

スキーマ照合

スキーマ照合の対象となるスキーマデータが入力されると，データから任意のカラム名ペアを取得し，ペア間の類似度を算出する．類似度の算出には名称そのものの類似性と，シノニムを組み合わせたハイブリッド型のマッチャーである Name Matcher[4]を使用する．

カラム名はトークンに分割し，トークンごとに 3-gram 単位での類似度および類義語辞書中の該当類義語同士の一致度に応じて類似度を算出し，総合してカラム名同士の類似度を算出する．

3. 評価実験

3.1 実験条件

実験にはデータベースのテーブルスキーマを用いた．本データは，エンジニア 33 名がデータベースの模擬設計を行い，資材発注システムおよび資材発注状況集計システムを想定して作成した．最低限必要な属性は提示した上で，テーブル分割，テーブル名およびカラム名は設計者の任意としている．データの一部を表１に示す．

表１．実験データの一部

テーブル名	カラム名
HATTYU	HATTYU_ID , HATTYU_YOSAN , HATTYU_NYUURYOKU , HATTYU_SYUBETSU , SHAIN_ID , HINMOKU_ID , HATTYU_SUURYOU , HATTYU_NOUNYUU , HATTYU_NOUNYUU_NENNDO , HATTYU_NOUNYUU_TSUKI , HATTYU_MITSUMORI , HATTYU_KUBUN , HATTYU_TANKA , KAISYA_ID
HINMOKU	HINMOKU_ID , HINMOKU_MEISYOU
SHAIN	SHAIN_ID , SYAIN_MEISYOU

ここで資材発注システムのスキーマをスキーマ A，資材発注状況集計システムのスキーマをスキーマ B とすると，それぞれのスキーマに作成者毎の連番を割り当て，スキーマ A の偶数番とスキーマ B の奇数番をデータセット１，スキーマ A の奇数番とスキーマ B の偶数番をデータセット２としてスキーマ照合を行った．

類義語辞書として，WordNet からノイズ無し（WSD 精度 100%）の語義として作成したもの，5%から 95%まで 5%刻みでランダムにノイズを加えて WSD を模擬したものを用いて評価を行った．ここで，ノイズ η %を加えた類義

† 三菱電機株式会社 情報総合技術研究所
Information Technology R&D Center,
Mitsubishi Electric Corporation

語辞書は，WSD の精度 $(100-\eta)$% での類義語辞書に相当する．なお，ノイズを加えた類義語辞書は乱数を変えて 5 種類ずつ評価した．

また，比較のため WSD を使わず，スキーマ中の出現形から WordNet を単純参照して得られる語義をすべて使って類義語辞書を生成した場合を評価した．日本語の WordNet 見出し語はあらかじめローマ字に変換することで，スキーマ中の語と対応を取っている．WSD としてみた場合，再現率 90.0%，適合率 10.6%であり，WSD の精度 19.0%（ノイズ 81.0%）に相当する．

最低限必要とした属性に関して，スキーマ A とスキーマ B 間の属性間の対応関係を人手により設定し，実験時の正解ペア（類似度の予測値が 1.0）として設定した．マッチャーにより算出した類似度は閾値を設けて閾値未満ならば 0.0，閾値以上ならば 1.0 とみなして正解ペアとの一致数を算出し，適合率と再現率の調和平均である F 値を算出しスキーマ照合の精度として評価した．閾値は 0.00 から 1.00 まで 0.01 刻みで設定し，最も良い F 値となる閾値を採用した．

以上のようにして，スキーマ照合結果が得られた全ペアである 218,940 ペアのうち，最低限必要とした属性同士のペアである 154,577 ペアを精度評価対象とした．本データの詳細を表 2 に示す．

表 2．実験データ

	全ペア	評価対象ペア	正解ペア
データセット 1	109,440	77,825	5,194
データセット 2	109,500	76,752	5,003

3.2 結果と考察

提案手法によりスキーマ A とスキーマ B のスキーマ照合を行い，算出された類似度と正解ラベルとの一致数を算出した．一定の WSD 精度が得られる場合，提案手法のほうが WordNet を WSD なしに類義語辞書として用いた場合よりも高い精度を得られた．例えば，類義語辞書のノイズ 0%の場合，F 値はデータセット 1 で 0.420，データセット 2 で 0.446 であり，WSD なしの類義語辞書を用いた場合の F 値 0.370 よりも良い結果となった．類義語辞書への WSD ノイズとスキーマ照合精度の関係を図 1 に示す．

ただし，類義語辞書のノイズ 0%と 95%の場合の精度の差は約 9%であり，WSD の精度がスキーマ照合の精度に与える影響は限定的であった．この原因としては，以下に考察するようにマッチャーのアルゴリズムが考えられる．

照合に失敗したペアでは，カラム名の命名において同じ単語が頻出しており，カラム名を構成するトークンの重要度を一律とし，類義語の登場回数で類似度を評価する NameMatcher では誤検出を起こしやすいと考えられる．

例えば，「発注番号」，「発注する品物の ID」，「発注者社員番号」という属性のカラム名が，それぞれ“HACCHUU_ID”，“HATTYU_HINMOKU_ID”，“HATTYU_SYAIN_ID”と命名されている場合，類義語「発注」が含まれ，「ID」という文字も一致していることから各ペア間の類似度はそれぞれ 0.67 となり類似度で区別がつかない．

また，スキーマ中に，WordNet の未登録語（“SYAIN(社員)”や“HATTYUSYA(発注者)”など）があり，類義語（上記の例では“EMPLOYEE”）との一致が判定できていないことも精度が向上しない要因と考えられる

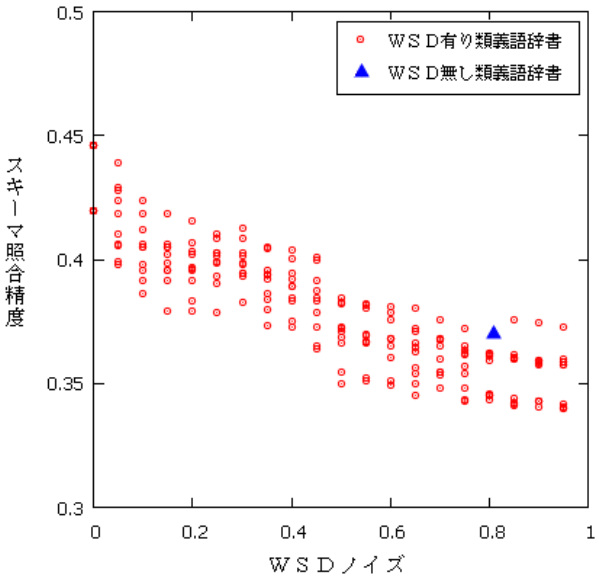

図 1．WSD ノイズとスキーマ照合精度

4．おわりに

本稿では，WSD の精度とスキーマ照合の精度の関係を測るための手法を提案した．既存のスキーマデータに提案手法を適用した結果，一定の WSD 精度が得られる場合，WSD なしの類義語辞書を利用するよりも高いスキーマ照合精度が得られることが示され，WSD の精度とスキーマ照合精度の関係が評価できたと言える．ただし，WSD の精度向上に対するスキーマ照合精度の向上は限定的であるため，手法を改善し，類義語辞書のノイズが精度に強く影響するケースで提案手法による評価を行いたい．

提案手法ではマッチャーに単純な照合を行う NameMatcher を採用したが，スキーマ照合を行う際に，頻出単語とそうでない単語の重み付けに偏りを持たせる tf-idf 重みを適用し，WSD に特化したマッチャーを用いることで精度の向上を図る．併せて，WordNet の未登録語に関して，特定のシステムに特化した用語リストを用いてカバーし，精度を向上させる方式を検討する

参考文献

[1]Rahm, E. and Bernstein, P.A. : A survey of approaches to automatic schema matching.VLDB J(10) pp.334-350,2001

[2] N. Ide and J. Veronis. “Introduction to the special issue on word sense disambiguation”. Computational Linguistics,Vol. 24, No. 1, pp. 1-40, 1998.

[3] R. Navigli. “Word Sense Disambiguation: a Survey”. ACM Computing Surveys, 41(2), ACM Press, pp. 1-69, 2009.

[4] Do, H.H. and Rahm,E. : COMA - A System for Flexible Combination of Schema MatchingApproaches,Proc. 28th Intl. Conference on Very Large Databases ,VLDB ,2002.