リンク予測を利用した語学学習用SNS上の友人関係グラフの分析

江原遥†, 宮尾　祐介*, 中川　裕志‡
† 日本学術振興会，* 国立情報学研究所，‡ 東京大学情報基盤センター
{ehara,yusuke}@nii.ac.jp, n3@dl.itc.u-tokyo.ac.jp

1　はじめに

グローバル化に伴い，近年では，英語とはじめとする第二言語（外国語）の習得は益々必須となり，第二言語を用いたコミュニケーションの量は言語研究の上で決して無視出来ない程にまで増大している．第二言語の学習用のプラットフォームとして，近年，Social Network Service (SNS) 上で学習者が相互に学び合うシステムである，**語学学習用 SNS** が注目されている．

語学学習用 SNS 上では，学習者が相互に影響しあって学習を行う**協調学習**が期待されている．個々の学習者の能力と協調の関係を明らかにしたい，という自然な動機がある．具体的にどのような協調があったかを知ることは難しい．そこで，本稿では，語学学習用 SNS 上で，ユーザが自分の意思に基づいて他のユーザとリンクする「友人関係リンク」に注目し，学習者の能力と友人関係リンクの関係について調査する．

学習者の「能力」については，様々な指標があり，それらを統一的に比較する事は難しい．そこで，本稿では，語学学習用 SNS のから入手可能なデータから，この能力を推定する事とする．具体的には，「誰がどの単語・フレーズを知っているか」の集合である**語彙知識データ**を用い，ここから，学習者の能力値を推定する．

語彙知識データから友人関係リンクを予測したいもう一つの理由は，プライバシー上の理由である．誰と友人関係にあるか，という友人関係リンクは，プライバシー上，保護が期待されるデータである．一方，例えば，「どの単語を知っているか」という語彙知識データについては，プライバシー上，どの程度保護されるべきなのか，その扱いがよく分かっていない．一方，語彙知識データは，読解支援の目的では有用である事が分かっている [1]．そのため，自分の語彙知識データを，サービス間で移動させたり，ダウンロードしたいというユーザが出てくる事は容易に考えられる．語彙知識データから友人関係リンクをどの程度予測できるか評価することは，新しいデータ資源として語彙知識データを扱う上で，重要な情報になると考えられる．

以上の理由から，本稿では，学習者の語彙知識データから，語学学習用 SNS のリンクを予測するタスクについて，予備的な実験を行い，その結果を報告する．

2　データ

本稿では，**smart.fm**[1] という語学学習用 SNS から，Web API を通じて公開されていたデータを，2010 年 7 月に収集したデータを用いる．smart.fm は，学習者が語やフレーズを学習していくための SNS で，最盛期には約 20 万人のユーザがいた．

smart.fm から，「あるユーザがある単語・フレーズを知っている/知らない」というデータ（以下，この形式のデータを語彙知識データと呼ぶ）を取得した．このデータがどのようにユーザから収集されるかについて，次に説明する．smart.fm には，いろいろな単語・フレーズ集である**コース**が用意されており，各ユーザは，ログインすると図 1 に示すような，単語／フレーズのリストを提示される．ユーザが既修のチェックを付けた語・フレーズを「知っている」，それ以外のユーザが学習している語を「知らない」として，二値に帰着させた．また，多くのユーザは，このシステムを少し使っただけでその後の使用をやめてしまい，このシステムの使用度に差がある．本稿では，1,000 語以上の語の学習歴が残されていた，3,281 ユーザだけを調査の対象とする．

次に，smart.fm の友人関係リンクについて説明する．本稿は，便宜上「リンク」という名称を用いるが，実際には，「ユーザ A はユーザ B を知っている」という関係を表す**有向グラフ**になっている．このような有向グラフの例としては，Web ページのハイパーリンクが挙げられる．smart.fm では，ユーザ A からユーザ B にリンクが出ている場合，ユーザ A はユーザ B の各コースの進捗状況などを閲覧することができる．これによって，「自分も頑張らなければ」といった競争心がユーザ A に生

[1] 現在は，**iKnow** という名前に戻っており，また，システムも大きく変わり，SNS としての機能は縮小されている．

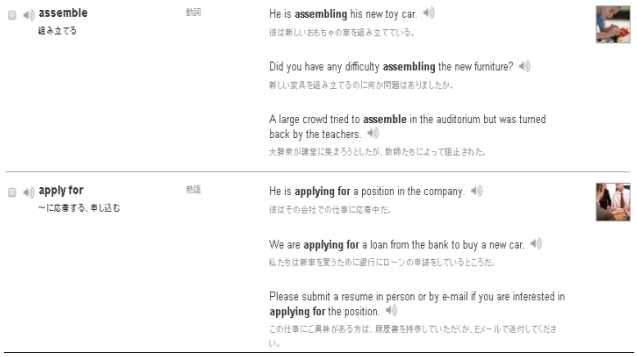

図 1: smart.fm のイメージ図．学習者は，単語・フレーズなどの項目の集合であるコースを選べる．各コースの中で，学習者は，既修の内容についてはチェックを付けられる．各コースは，各項目について，例文の穴を埋めたり，正しい意味を選択したりするなどの問題を練習していく形式で進められる．この時，既修のチェックが付けられた項目については飛ばされる．厳密には，この図は，2013 年に後続サービスの iKnow (`http://iknow.jp/`) から取得・引用した．

まれ，ユーザ同士が互いに高め合う事が意図されていると思われる．本稿では，上記の 3,281 ユーザのみをノードとして残し，これらのユーザ間のグラフ構造のみを対象とした結果，3,281 ノード，27,786 リンクからなるグラフが得られた．

3 Rasch モデル

本節では，語に対する被験者の反応の分析・予測に広く使われる，項目反応理論の一般的なモデルである，Rasch モデル [2] を説明する．Rasch モデルでは学習者 u が語 v を知っている確率を次の式でモデル化している．

$$P(y=1|u,v) = \sigma(a_u - d_v). \quad (1)$$

ここで，$\sigma(t) = (1+\exp(-t))^{-1}$ は，ロジスティックシグモイド関数である．

式 (1) には，データから推定されるパラメタが 2 種類ある． 1 つは，ユーザ u の能力値パラメタ a_u，もう 1 つは，単語 v の難易度パラメタ d_v である．

図 3 に，ユーザの能力値 a_u のヒストグラムを示す．図 3 から分かるように，全体的に，能力値は釣鐘状の分布をするのが普通であり，今回のデータに対しても能力値はこのような分布になっている．ただし，図 3 の左側が欠けており，能力値の極端に低いユーザは欠けている．これにはいくつかの解釈が可能である．まず，語学学習用 SNS を使って英語力を向上させようというユーザは，そもそも，ある程度の英語力を持っていると考える事が可能である．ユーザが「英語力を向上させたい」と思う

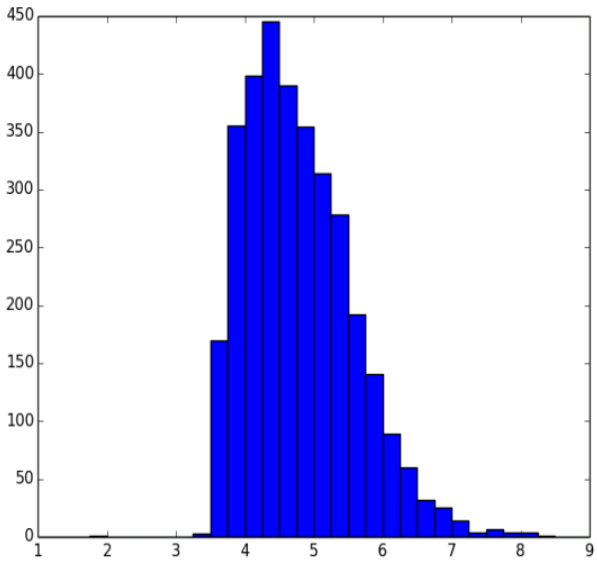

図 2: 能力値のヒストグラム．横軸は能力値，縦軸は人数．

	Dst. 低	Dst. 高
Src. 低	35	74
Src. 高	81	166

表 1: 能力値と友人リンクの傾向．高 : 能力値上位 200 人，低 : 能力値下位 200 人．各マスの数値はリンクの数，Src. はリンク元，Dst. はリンク先を表す．例えば，Src. 低と Dst. 高のマスは，低能力群から高能力群へ張られたリンクの数を表す．

程度には，英語の能力が必要とされる環境にいると考えられるからである．また，本稿で扱うデータは，ユーザの能力値が評価可能なように，一定以上の学習歴のあるユーザに限定している事も関連していると思われる．

表 1 に，能力値と友人関係リンクの傾向を示す．高能力群として，能力値 a_u の上位 200 人，低能力群として，能力値 a_u の下位 200 人を選び，その間の友人関係リンクの頻度を示す．高能力群が，低能力群よりも有意に他からリンクされる傾向が高い事が分かる（χ 二乗検定で有意，$p < 0.05$）．

4 手法

本稿では，[6] を基に，リンク予測を二値分類問題に帰着させて解く．図 3 に二値分類問題への帰着方法を図示する．ユーザ A とユーザ B の素性ベクトルが与えられた時に，ユーザ A とユーザ B の間に，A → B へのリンクがあるかないかを予測する．

予測にあたっては，確率値が出ることと，理論的な妥当性から，ロジスティック回帰を用いる [6]．単純にユー

ザ A とユーザ B の素性ベクトルをそのまま繋げて，ロジスティック回帰の識別器に流すことも考えられるが，リンク予測の場合，**組み合わせ素性**が重要である事が知られている．例えば，ユーザ A がある単語を知らず，ユーザ B がある単語を知っていた時にだけ，友人関係のリンクが張られる，という場合がこれに該当する．

このような組み合わせ素性は，クロネッカー積の**混合積性質**を利用したカーネルを用いて，カーネルを通じてロジスティック回帰に導入することが可能である事が知られている [6]．本稿では，比較のため，カーネルロジスティック回帰で最適なパラメタを求めるために，カーネル行列を丸ごと陽に持ち，Iterative Reweighted Least Squares(IRLS) 法 [5] を用いた独自の実装で解いている．

5 実験

リンク予測タスクでは，グラフによって，そもそもリンクの頻度が違うので，単純に精度で性能を評価すると，グラフによってベースラインが異なってしまう．これを防ぐため，リンク予測の評価には，通常，Area Under the Curve (AUC) が用いられる [6] ので，本稿でも，AUC を用いて評価した．AUC は，Receiver Operating Characteristic (ROC) 曲線の下部の面積である．ROC 曲線は，横軸に True Positive Rate (TPR)，縦軸に False Negative Rate (FNR) を取って，システムの性能を表示したものである．TPR,FPR は次のように定義される．

$$\text{TPR} = \frac{\text{うち，実際の正例件数}}{\text{システムが正例と回答した数}} \quad (2)$$

$$\text{FNR} = \frac{\text{うち，実際には正例の件数}}{\text{システムが負例と回答した数}} \quad (3)$$

システムは何らかのスコアを返すとする．例えば，今回のタスクのロジスティック回帰であれば，「学習者間にリンクがある確率」をスコアとみなせる．通常はこの確率が閾値 0.5 以上の時システムは「リンクあり」と回答するが，ROC 曲線では，この閾値 0.5 を変化させる事によってシステムの回答を変化させ，複数の TPR,FPR の点をプロットしていく．例えば，ROC 曲線で一番下の点は，システムが全てのテストデータに対して「リンクなし」と返した場合に対応する．AUC は 0 から 1 の値を取り，大きい方が良い．システムの返すスコアの降順にデータを並べた時に，綺麗に，上半分が正答，下半分が誤答の形になっていれば，AUC は最高の 1.0 の値を取る．ベースライン（最も悪い場合）[2]は，スコアに関係なくランダムにデータを並べた場合であり，ROC 曲線では左下の点と右上の点をつないだ直線で表され，AUC は 0.5 の値を取る．

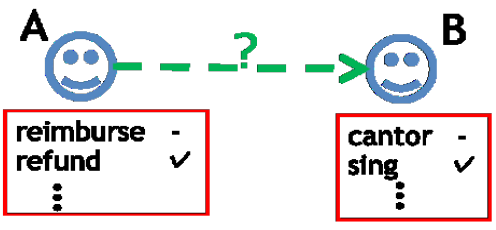

図 3: リンク予測の図式化．二人の素性ベクトル（赤枠）が与えられた時に，A から B へのリンクがあるかないかを予測する二値分類問題になっている．

リンクの有無を予測する場合（リンクがある場合を正例），リンクの方向を予測する場合の 2 種類の分けて実験を行った．訓練データ・テストデータは，グラフからランダムにノードを取って来て，正例・負例の数が同じになるように調整した．訓練データは 3,000 件（すなわち，ユーザペア 3,000 ペア）の場合と 10,000 件の場合を試した．素性は，§3 で示した，Rasch モデルから取った能力値のみを素性に用いた**能力値のみ**と，ユーザペアのそれぞれの語彙知識をそのまま素性に使った**全語彙知識**，さらに，語彙知識間の組み合わせ素性を，カーネルロジスティック回帰を用いて入れた場合**全語彙知識＋組み合わせ素性**を試した．

表 2 に実験結果を示す．AUC を用いているので，ベースラインは，0.5 である．まず，能力値のみを用いた場合でも，ベースラインよりわずかに AUC が高い事が分かる．実際，AUC ではなく，対応する精度を用いて検定を行った所，ベースラインに対して，能力値の精度は有意であった．また，表 2 より，リンクの有無に関しては，組み合わせ素性を用いても用いなくても，性能に大差はないことが分かる．一方，リンクの方向に関しては，組み合わせ素性を用いることにより，AUC が大きく向上することが分かる．

図 4 に ROC 曲線を示す．標準的な ROC 曲線の形になっていることが分かる．また，左側の立ち上がりが若干早く，右の方で TPR が 1.0 にゆっくり近づく傾向が分かる．

6 おわりに

本稿では，語学学習用 SNS 上の友人関係リンクを，語学能力パラメタから予測するタスクを提案し，このタスクに対する予備的な実験・評価を行った．今回のグラフでは，リンクは有向と考えている．評価実験の結果，リンクの方向の予測に対しては，組み合わせ素性が有向で

[2]ベースラインを下回るシステムは，システムが回答する正例・負例を反転させることで，ベースラインを上回るシステムとみなせる．

	リンクの有無		リンクの方向	
	訓練 3K	訓練 10K	訓練 3K	訓練 10K
ベースライン	0.5	-	0.5	-
能力値のみ	0.582	-	0.579	-
全語彙知識	0.770	0.827	0.520	0.508
全語彙知識+組み合わせ素性	0.757	0.823	0.531	0.631

表 2: 実験結果．セル中の数値は全て AUC．リンクの有無は，リンクの有無を予測した場合，リンクの方向は，リンクの方向を予測した場合の値．訓練 3K：訓練例 3,000 ペア．訓練 10K：訓練例 10,000 ペア．テストは 2,000 ペアで行った．

あるが，リンクの存在の有無に対しては，組み合わせ素性を入れた場合と入れない場合で性能（AUC）に大きな違いがなかった．

将来の課題としては，次のようなものが挙げられる．まず，本研究は分析が主目的である．通常のロジスティク回帰であれば，学習後のモデルに対して，素性に対応する重みを実際に確認し，有効な組み合わせ素性を抽出してみることによって，分析の幅を広げる手法が利用できる．しかし，今回はカーネルロジスティック回帰を利用してしまったため，この手法はナイーブには適用できない．カーネルから重みベクトルを復元することは可能であるが，この計算は，ナイーブには素性数の 2 乗のオーダであり，高速化が必要である．そこで，著者らは，現在，カーネルから復元された重みのうち，大きい方から K 個までを抽出する手法を考案中である．実データに対して用いた予備実験では，ナイーブな場合に比べて相当の高速化が可能なようである．カーネルロジスティック回帰は，今回のタスクに限らず，広くテキスト間の関係を予測する目的で利用できると考えられるので，より幅を広げた自然言語処理のタスクに対して，この高速化手法を適用し，復元された重みを分析する事が，今後の重要な課題である．

また，今回は，カーネルロジスティック回帰を古典的な IRLS 法 [5] で解いたが，この手法は，データ数が多くなった時に必要とする空間計算量も時間計算量も大きいという問題がある．カーネルロジスティク回帰に対しては，近年，[3, 4] といった，スパースでオンラインな学習法が提案されているため，これらを利用し，元から疎な構造が内在すると思われる実データに対しては，高速に動作するようにすることも，今後の課題の 1 つである．

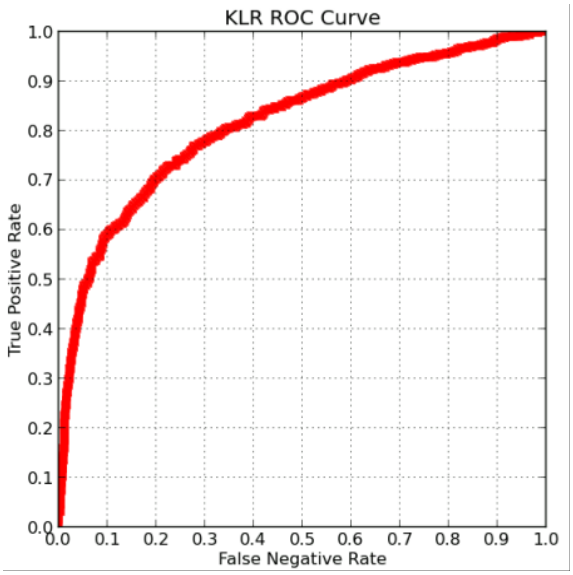

図 4: ROC 曲線．表 2 の訓練 10K で全語彙知識+組み合わせ素性でリンクの有無を予測した場合．

参考文献

[1] Y. Ehara, N. Shimizu, T. Ninomiya, and H. Nakagawa. Personalized reading support for second-language web documents. *ACM Transactions on Intelligent Systems and Technology*, 4(2), 2013.

[2] G. Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests.* Danish Institute for Educational Research, Copenhagen, 1960.

[3] L. Zhang, R. Jin, C. Chen, J. Bu, and X. He. Efficient online learning for large-scale sparse kernel logistic regression. In *AAAI-12*, 2012.

[4] L. Zhang, J. Yi, R. Jin, M. Lin, and X. He. Online kernel learning with a near optimal sparsity bound. In *ICML 2013*, 2013.

[5] J. Zhu and T. Hastie. Kernel logistic regression and the import vector machine. In *NIPS*, pp. 1081–1088, 2001.

[6] 鹿島． ネットワーク構造予測． 人工知能学会誌, 22(3):344–351, 2007.