COLING2014 参加報告（その 3）
– 分野適応と知識獲得の論文紹介 –

斉藤いつみ[†]

1　はじめに

本稿では，2014 年 8 月 23 日〜29 日にアイルランド・ダブリンで開催された Coling2014（The 25th International Conference on Computational Linguistics）の参加報告を行う．会議の概要については他の参加者より報告があるため，本稿では特に，テキスト解析の新しい分野への適用と，知識獲得に関する文献について報告を行う．本稿では概要を記述するため，詳細については本論文を参照されたい．

2　テキスト解析の新しい分野への適応

2.1　テキスト正規化と翻訳：A Framework for Translating SMS Message

この論文 (Rangarajan Sridhar, Chen, Bangalore, and Shacham 2014) では，現在コミュニケーションツールとして広く用いられているショートメッセージサービス（SMS）テキストの翻訳（英語とスペイン語間の翻訳）を試みている．多言語コミュニケーションを考える際に SMS のようなテキストを翻訳できることは有用であると考えられるが，SMS テキストの翻訳においては (1) データの入手困難性，(2) ノイズ（省略，スペルエラー，句読点の欠落等）のあるテキストの扱い，といった大きな問題がある．著者らは，(1) の問題に対して SMS の代替となりうる 3 つのデータ（speech-based SMS, Amazon Mechanical Turk, Twitter）を使ってモデルを学習する方法を提案している．(2) の問題に対しては，SMS 翻訳のタスクを大きく 3 つの処理（テキスト正規化，フレーズセグメンテーション，機械翻訳）のパイプライン処理として設計することで精度を向上させている．テキスト正規化についてはここ数年で多くの研究が発表されているが，これらが教師あり学習に基づくものが多いのに対し，著者らは教師なしのテキスト正規化手法を提案している．

評価は，独自の翻訳 SMS サービス（英語-スペイン語間）を構築して客観・主観評価実験を

[†]NTT メディアインテリジェンス研究所, NTT Media Intelligence Laboratories

Canonical form	Noisy form
love	loveeee, loveeeee, looove, love, wuv, wove
starbucks	starbs, sbucks
tomorrow	tmrw, tomorrow, 2moro, tmrrw, tomarrow

表 1　獲得した正規語と崩れ語の例（Rangarajan et al(2014) より抜粋)

行った．その結果 BLUE スコアで 31.25 (英語-スペイン語)，37.19 (スペイン語-英語)，主観評価においては適切性（adequacy）に関して大半の被験者が十分であると評価し，流暢さ（fluency）は good から non-native の間（5 段階で 3〜4 点）という比較的良好な結果を得た．下記に，テキスト正規化とフレーズセグメンテーションについて概要を記述する．

Unsupervised SMS normalization

著者らは正規語と崩れ語の対応づけ辞書を獲得するため，Twitter テキストを用いて単語の分布表現を学習し，コサイン類似度と文字列類似度の双方の尺度において各単語の近傍に属する単語を獲得している．英語とスペイン語の双方で行っているが，いずれも 1best で評価した際に高い精度で正規語と対応する崩れ語の抽出ができた．抽出した英語の対応づけ例を表 1 に示す．なお，今回の実験では 1 単語対 1 単語の対応付けを学習しているが，many to one や many to many の対応づけに拡張することも可能としている．

Phrase segmentation

SMS では，多くの文章で句読点が欠落している．例えば，“hi babe hope you’re well sorry i missed your call” という文は “hi babe. hope you’re well. sorry, i missed your call.” と解釈されるべきである．このように欠落した句読点の位置を自動的に推定するフレーズセグメンテーションモデルを推定し，翻訳の前処理として用いる．このセグメンテーションを行わない場合，翻訳の精度は大きく低下することが実験より明らかになった．

2.2　分野適応 1 : Adapting taggers to Twitter with not-so-distant supervision

この論文 (Plank, Hovy, McDonald, and Søgaard 2014) は，Twitter テキストを対象として POS tagging（品詞タグ付け）と NER（固有表現抽出）についてドメイン適応を行い提案手法によって従来モデルに比べ POS tagging で 8%，NER で 10%のエラー軽減を達成した，という論文である．近年 Twitter を用いた解析が増えているが，テキストから自動的に情報抽出を行うためには基本的な解析（例えば POS tagging や NER）が精度良くできることが求められる．一方 Twitter は従来のテキストに比べ，スペリングの多様性，省略表現の多用，崩れた文法などが存在し，既存の解析器では正しく解析することが難しい．

ツイート本文（リンク付き）	#Localization #job: Supplier / Project Manager - Localisation Vendor - NY, NY, United States **http://bit.ly/16KigBg** #nlppeople
リンク先文	The Supplier/Project Manager performs the selection and maintenance . . .

表 2　ツイート本文とリンク先文の例（Plank et al(2014) より抜粋)

既存研究では Twitter テキストに対し，比較的少ないラベル付きデータを用いて POS tagging の精度を向上する手法が提案されているが，これらのモデルは対象ドメイン（Twitter）にオーバーフィットする傾向があり，対象ドメイン以外のテキスト（例えば新聞など）では解析精度が悪くなる場合も生じるという問題があった．

著者らはこれらのモデルバイアスを軽減するため，少ないラベル付きデータと大量のラベルなしデータを用いて，より頑健なドメイン適応を行う手法を提案している．このような手法自体は，self-training（自己学習）や Distant supervision（弱教師あり学習）と呼ばれる枠組みでこれまでにも研究が行われてきたが，従来の Distant supervision で主に使われていた Wiktionary（Wikipedia の辞書版のようなもの）に加え，新たなリソースとしてツイート文のリンク先テキストを利用するというのがこの研究のポイントである．表 2 に論文中で示されているリンク付きツイートの例を示す．リンク先文とツイート本文は多くの単語を共有しているが，リンク先文の方が文脈の中で単語が現れていることがわかる．例えば，表 2 の 2 つの文章（ツイート本文とリンク先文）を既存の POS tagging モデルで自動解析した場合，ツイート本文の “Supplier” は誤って “形容詞” と解析されるが，リンク先文の “Supplier” は正しく “名詞” と解析される．著者らの着目点は，上記のようにリンク先文の解析結果の方が信頼度が高いことを利用し，ツイート本文のラベル推定を行う際に，リンク先文のラベル推定結果を反映して推定する，という点である．実験では，リンク先文の結果を反映する場合と，Wiktionary のラベルを反映する場合の組み合わせも実験しており，何も情報を反映しない場合（ツイート本文の自動解析結果をそのまま用いる場合）と比べて双方の情報を組み合わせて反映する場合が最も精度がよかったことを報告している．

2.3　分野適応 2 : Automatic Corpus Expansion for Chinese Word Segmentation by Exploiting the Redundancy of Web Information

この論文 (Qiu, Huang, and Huang 2014) も発想としては (Plank et al. 2014) と非常に似ており，ラベルなしコーパスを学習データとして追加し，中国語単語分割モデルの改善を行う話である．この研究では，既存モデルで解析が難しい文の代替として，関連する語を含んだより解析しやすい文を web 上から探してきて学習データに追加する，というアイデアを用いている．ここで，“解析が難しい文” とは例えば次の文 (a) のような文である（以下例文は論文中よ

り抜粋)．(a)“欧莱雅美宝 (L’Oreal, Maybelline)” この文において，“欧莱雅 (L’Oreal)”と “美宝 (Maybelline)”の双方が未知語である場合正しく単語分割するのは難しい．しかし，関連する文である (b)“我使用美宝 (I use Maybelline)”や (c)“欧莱雅的品 (production of L’Oreal)”は (a) に比べて単語分割が容易であることが推測できる．さらに文 (b) や (c) のように，元文 (a) と関連していて，(a) に比べ解析しやすい文は大量のテキストが存在する web 上から探すことができる．このことから，(a) のような文に対し，解析しやすい (b) や (c) の文を抽出し，それらの自動解析結果を正解データに追加しよう，というのが著者らの基本的なアイデアである．

上記のアイデアに基づき，この論文では学習データとして追加するテキストを次の三段階の方法で選択している．まず 1 ステップ目として，ラベルありデータを用いて学習したベースモデルで対象ドメインテキストを解析し，解析の信頼度が閾値以下の（既存モデルで解析が難しいと考えられる）テキストを抽出する．次に 2 ステップ目として，抽出したテキストの部分文字列をシードとして web から対象テキストと関連度が高いテキストを抽出する．3 ステップ目に，抽出した関連度の高いテキストを自動解析し解析の信頼度が閾値以上の候補のみを学習データとして追加する．というフレームワークである．この際，信頼度の基準はいくつかの指標を用いて独自に設定している．新しい学習データから更新されたモデルを用いて上記のプロセスを繰り返しモデルを更新していく．

Twitter のような新しい分野のテキストからの情報抽出やマイニングといったタスクが増えると同時に，これらのテキストを頑健に解析しようとする研究はここ数年で増えてきている．特に，2.1 で紹介した論文のように，崩れたテキストを正規化し既知語にマッピングして解析しようとする研究 (Han and Baldwin 2011; Li and Liu 2012; Liu, Weng, and Jiang 2012) と，2.2, 2.3 で紹介した論文のようにラベルなしデータを学習データとして追加したり，外部リソースを用いたりしてモデルの改善を行う研究 (Mintz, Bills, Snow, and Jurafsky 2009; Jiang, Sun, Lü, Yang, and Liu 2013) はいずれもここ数年の国際会議でもよく目にするテーマであり，新たなドメインのテキストに対する解析手法への関心が高まっていることが伺える．また今回紹介した論文はいずれもシンプルな手法であり，実用的な適用しやすさも重視されていることを感じた．

3　知識獲得：Triple based Background Knowledge Ranking for Document Enrichment

この論文 (Zhang, Qin, Liu, and Zheng 2014) では，対象テキストと関連度の高い背景知識を外部リソースから獲得するタスクに取り組んでいる．テキスト中では，一般的でよく知られている背景知識は省略して書かれることが多い．例えば，次の 2 つの文章を考えよう．（いずれも論文中より抜粋） S1: Coalition may never know if **Iraqi** president Saddam Hussein survived a U.S. air strike yesterday. S2: A B-1 bomber dropped four 2,000-pound bombs on a building

in a residential area of **Baghdad**. これらの文において，“Baghdad が Iraqi の首都である”という背景知識は明示されないことが多い．人間はこのような背景に仮定される一般的な知識を文章を読みながら補完し，文間のつながりや意味を理解するが，機械が処理を行う場合には背景知識の欠落が意味理解の大きな障害となる．上述のような一般的な背景知識を大規模な外部リソースから獲得し，テキスト解析に活かそうとする試みは多く存在する．例えば情報検索や共参照解析，文書分類，エンティティの曖昧性解消などのタスクで効果があることが知られている (Pantel and Fuxman 2011; Volha, Claudio, Luciano, and Kateryna 2010)．

従来は，背景知識の獲得源として Wikipedia とオントロジーが主要なリソースであったが，Wikipedia はカバレッジは高いものの対象テキストと関連のない余分な情報も多く含まれノイズが大きいこと，オントロジーは精度はよいがカバレッジが低いことが課題として残っていた．著者らは，($argument_1$, $predicate$, $argument_2$) の三つ組を知識獲得の対象とし，外部リソースから得られた知識に対し，対象テキストとの関連度に基づいたランキンキングモデルを適用することでカバレッジと精度の双方を満たす知識獲得を行うことを提案している．基本的なアイデアは，対象テキストと外部リソースから ($argument_1$, $predicate$, $argument_2$) の三つ組を抽出し，これらをノードとして，対象テキストと外部リソースのノードの関係性をグラフで表現することである．このグラフを用いて対象テキストと背景知識の意味的な関連性をグラフ伝搬の手法によって計算しランキング評価した結果，ランキングの代表的な評価指標 MAP と $P\&Q$ において，ベースラインを上回る結果が得られた．

まとめで著者らも述べているように，高度な意味処理に基づく NLP の応用分野（一貫性の評価や質問応答など）は重要度を増してきている．その中で，対象テキスト中に現れない一般知識や背景知識をどのように獲得し，それらをどのように構造化して保持するかという問題は，今後も引き続き重要な問題となることが予想される．

4　おわりに

本稿では，国際会議 Coling2014 の発表論文について紹介した．筆者は自然言語処理の国際会議に参加したのが今回初めてであったが，幅広い分野からの発表があったことが印象的で言語処理研究の分野としての活気を感じた．上記で紹介した論文の他にも，木構造カーネルに単語の分布表現を導入する論文 (Ferrone and Zanzotto 2014) や，暗号解読の問題で単語レベルの言語モデルと文字レベルの言語モデルを組み合わせてモデル化し，モンテカルロ木探索によって良い解を探索するという論文 (Hauer, Hayward, and Kondrak 2014) も興味深かった．

また会議を通して，意味の分布表現（distributional semantics）や構成意味論（compositional semantics）に関する研究が多く見られた．Best paper が単語レベル・文レベルの意味の分布表現を用いた名詞の関係分類の論文 (Zeng, Liu, Lai, Zhou, and Zhao 2014) であったことや，チュー

トリアルで word2vec の提案者である Mikolov 氏が講演されていたことからも，単語から句，文の意味の分布表現とその応用といったテーマに大きな期待と関心が寄せられていることを感じた．その他には認知言語学の観点からノイジーチャネルモデルと実験に基づいて言語の語順に関する考察を展開した Gibson 氏による招待講演 (Gibson 2014) も興味深かった．次回は日本（大阪）で開催されるため，引き続き日本からも多くの論文が発表されることを期待したい．

謝辞

本稿の一部は 2014 年 9 月に開催された NLP 若手の会シンポジウムにおける Coling2014 参加報告に基づく．シンポジウムの運営に携わられた方々，コメントを下さった方に感謝申し上げる．また，NTT メディアインテリジェンス研究所の今村賢治氏，東中竜一郎氏，西川仁氏から情報共有・アドバイスをしていただいた．ここに感謝の意を記す．

参考文献

Ferrone, L. and Zanzotto, F. M. (2014). “Towards Syntax-aware Compositional Distributional Semantic Models.” In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 721–730. Dublin City University and Association for Computational Linguistics.

Gibson, E. (2014). “Language for Communication: Language as Rational Inference.” In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 781–782. Dublin City University and Association for Computational Linguistics.

Han, B. and Baldwin, T. (2011). “Lexical Normalisation of Short Text Messages: Makn Sens a #Twitter.” In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pp. 368–378.

Hauer, B., Hayward, R., and Kondrak, G. (2014). “Solving Substitution Ciphers with Combined Language Models.” In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 2314–2325. Dublin City University and Association for Computational Linguistics.

Jiang, W., Sun, M., Lü, Y., Yang, Y., and Liu, Q. (2013). “Discriminative Learning with Natural Annotations: Word Segmentation as a Case Study.” In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 761–769 Sofia, Bulgaria. Association for Computational Linguistics.

Li, C. and Liu, Y. (2012). "Improving Text Normalization using Character-Blocks Based Models and System Combination." In *Proceedings of COLING 2012*, pp. 1587–1602.

Liu, F., Weng, F., and Jiang, X. (2012). "A Broad-Coverage Normalization System for Social Media Language." In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1035–1044.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). "Distant supervision for relation extraction without labeled data." In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 1003–1011. Association for Computational Linguistics.

Pantel, P. and Fuxman, A. (2011). "Jigs and Lures: Associating Web Queries with Structured Entities." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 83–92. Association for Computational Linguistics.

Plank, B., Hovy, D., McDonald, R., and Søgaard, A. (2014). "Adapting taggers to Twitter with not-so-distant supervision." In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1783–1792.

Qiu, X., Huang, C., and Huang, X. (2014). "Automatic Corpus Expansion for Chinese Word Segmentation by Exploiting the Redundancy of Web Information." In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1154–1164.

Rangarajan Sridhar, V. K., Chen, J., Bangalore, S., and Shacham, R. (2014). "A Framework for Translating SMS Messages." In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 974–983.

Volha, B., Claudio, G., Luciano, S., and Kateryna, T. (2010). "Using Background Knowledge to Support Coreference Resolution." In *ECAI 2010 - 19th European Conference on Artificial Intelligence, Lisbon, Portugal, August 16-20, 2010, Proceedings*, pp. 759–764.

Zeng, D., Liu, K., Lai, S., Zhou, G., and Zhao, J. (2014). "Relation Classification via Convolutional Deep Neural Network." In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 2335–2344. Dublin City University and Association for Computational Linguistics.

Zhang, M., Qin, B., Liu, T., and Zheng, M. (2014). "Triple based Background Knowledge Ranking for Document Enrichment." In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 917–927.

略歴

斉藤いつみ（正会員）：2010 年早稲田大学理工学部卒業．2012 年東京大学大学院工学系研究科都市工学専攻博士前期課程修了．同年，NTT（メディアインテリジェンス研究所）に入社．自然言語処理の研究に従事．

（20xx 年 x 月 xx 日依頼）
（20xx 年 x 月 xx 日受付）