

横山研 競プロゼミ

木山 朔

自己紹介

名前 木山 朔 (きやま はじめ)

所属 東京都立大学 M1

専門 自然言語処理 (MT, GEC)

出身 神奈川県相模原市

Twitter @rudo_halo

Atcoder 茶色(Highest 681)

コメント

競プロゼミですが…

学会参加の感想をメインに語ります



参加の動機

1. ポスター発表を通じて自分の研究をどうするか決める
2. 懇親会などで知り合いを増やす

参加時の先生からのコメント

学会参加の一番の目的は
学外の人との交流です

スケジュール

日	月	火	水	木	金	土
12	13	14	15	16	17	18
GMT+09 NLP2023 @ 沖縄						
午前8時						
午前9時						
午前10時	羽田空港 午前10時～11時		機械学習 午前9:30～11:15	言葉の評価と品質推定 午前9:30～11:15	ポスター発表 午前9:15～11:15	WS：深層学習時代の計算言語学 午前9:30～午後12時
午前11時	搭乗手続き 午前11時～午後12時		形態素解析 午前11:30～午後1時	ポスター・移動 午前11:30～午後1時	meeting・気絶 午前11:30～午後1:15	
午後12時	羽田空港から那覇空港 午後12時～3時		ChatGPT 午後1時～2時	NLP B4ランチ会 午後1時～3:30	突撃会 午後12時～1:30	那覇空港から羽田空港 午後12時～2:30
午後1時		チュートリアル2：NLPからみた古典語と現代語 午後1時～2:30	埋め込み表現1 午後2:15～3:45		形式言語・マルチモーダル 午後2:15～4時	
午後2時		チュートリアル4：構文文法の基本的な考え方 午後2:45～4:15	埋め込み表現2 午後4時～5:45	招待講演：記号創発 午後4:15～5:15	招待講演：危機言語 午後4:15～5:15	
午後3時	ホテル移動 to 県庁前 午後3時～4時		含意・言い換え 午後6時～7:30	ポスター 午後5:30～7時	クロージング 午後5:30～7時	
午後4時			参加者交流イベント 午後7:30～8:30	B4夕飯会 午後7時～9時		
午後5時						
午後6時						
午後7時						
午後8時						
午後9時	ちゅらデータ懇親会 午後8:30～11:30		参加者交流イベント二次会 午後9時～午前1時		都立大+名大+a 懇親会 午後8:30～午前12時	オンライン懇親会も 参加しました 3/19
午後10時						
午後11時						

参加の動機

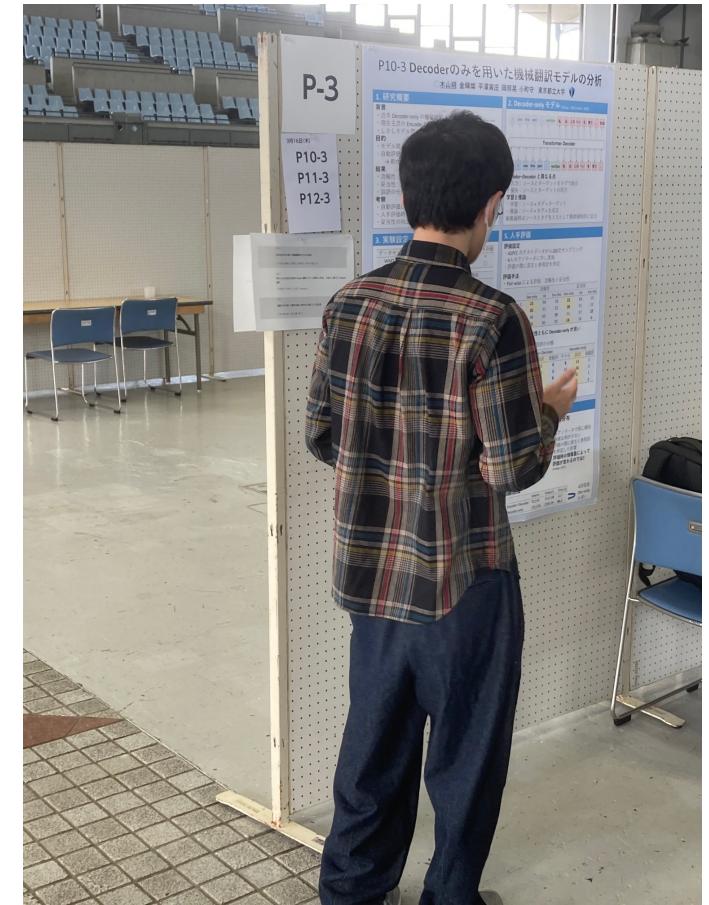
1. ポスター発表を通じて自分の研究をどうするか決める
2. 懇親会などで知り合いを増やす

現地の雰囲気

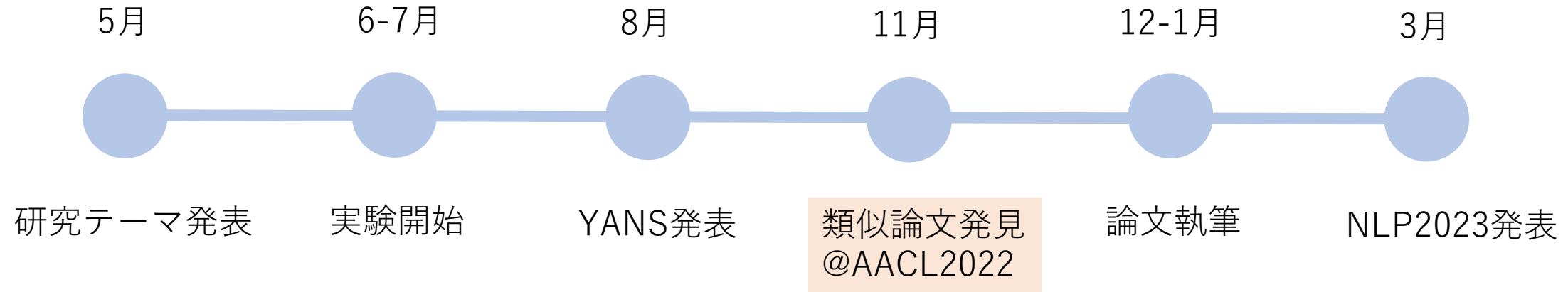
- とにかく人が多い！
 - ポスターセッションは本当に多い
 - 声が通らない
 - 口頭セッションも多い
 - 後ろの席だとスライドが見えない
 - オンラインと対面のずれを認識
- スポンサーブースでいろんな企画
 - 物もらえるの多かった印象
 - 全ては回れなかった…
 - 事前にどこ巡るか準備した方が良い

ポスター発表

- P10-3 「Decoderのみを用いた機械翻訳モデルの分析」
 - 3/16（木） 9:40-11:10
 - 20分前にポスターを設営したら人が集まった
 - 発表している時に共著者の方から水を差し入れ
 - 終了後も質問を受けたりした
 - **終わった後は気絶**, meeting
 - 研究テーマは変更の方針
- 詳しい資料はHP上に公開しています
 - ポスターとスライド両方用意しています
 - このあと話しかけて聞いても大丈夫です



研究のこぼれ話



- 類似論文が国際会議に出現
- 手法ではなく分析メインの研究として続ける
- 実験設定に妥当でない部分が多くかった…
- ➔ 研究の速さを実感, 実験デザインが大事

参加の動機

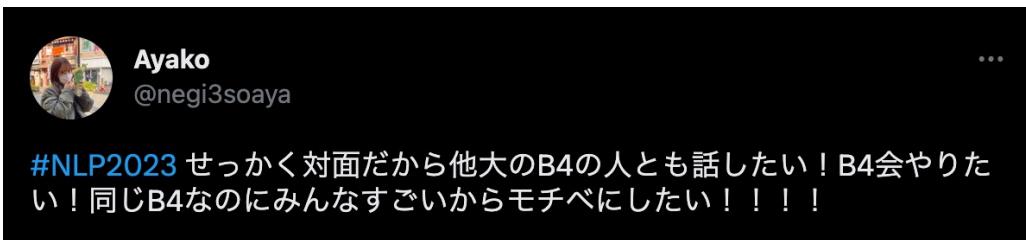
1. ポスター発表を通じて自分の研究をどうするか決める
2. 懇親会などで知り合いを増やす

参加者交流イベント二次会

- 先着40名の枠に参加できた！
- 別の年代の方と交流
 - 誰と一緒に研究するかが重要
 - アメリカ留学はいいぞ
 - 社会人Dはいいぞ
 - などなど
- 自分の発表を見に来てくれた人も
 - うれしい

B4会

- 同期との雑談から発足
 - 「B4同士なら比較的話しやすいし知り合いになれるそうじゃない？」
- 自分はslackの連絡を担当
- 参加者は40人超え！
- 当日はかなり緊張した…
- 有志で遊びに行った人たちも



競プロ的な知識が使えそうな論文1

最小コスト法に基づく形態素解析における CPU キャッシュの効率化

- メモリのランダムアクセスに着目

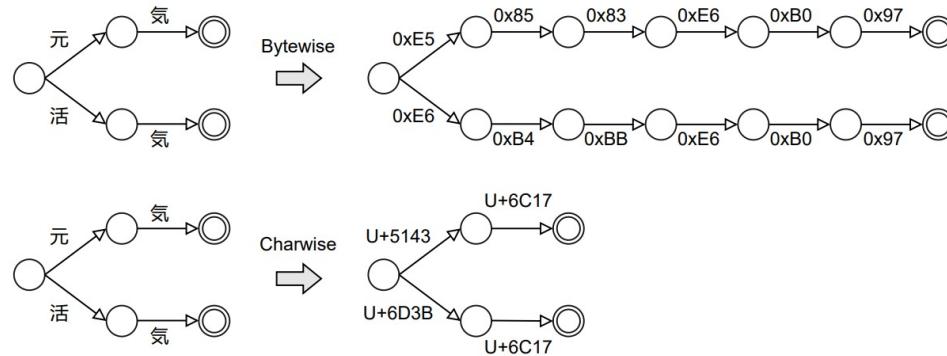


図 2 形態素「元気」「活気」を登録したトライ。文字列のエンコードが Unicode である場合の、上図は Bytewise で表現した例、下図は Charwise で表現した例を示す。

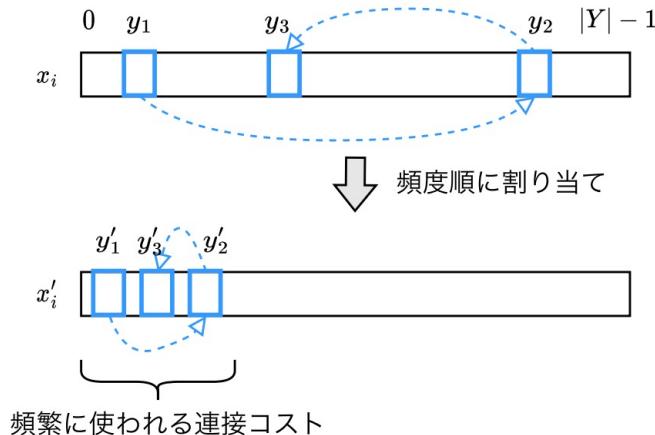


図 3 文脈 ID を頻度順に割り当てる例。ある左文脈 ID x_i に対応する行と、 x_i について右文脈 ID y_1, y_2, y_3 の順でコスト値を参照した様子を図示している。 x'_i, y'_j は頻度順に割り当てられた文脈 ID を表す。 y_1, y_2, y_3 が頻繁に使用される文脈 ID であった場合、 y'_1, y'_2, y'_3 は小さい値となる。つまり、その連接コスト値は行の先頭付近に密集し、参照の局所性は改善する。

競プロ的な知識が使えそうな論文2

最長一致パターンに基づく高速・高精度な日本語形態素解析

単純に深層学習時代にルールベース的な手法が提案しているのが面白い

精度をそのまま、速度向上を目指している

概要

膨大な量のテキストを解析したり、言語処理応用で大量にユーザのクエリを処理する場合、処理効率の悪いモデルは高精度でも利用し難い。本稿では高効率な手法の精度を改善すべく、最長一致パターンに基づく高精度な形態素解析手法を提案する。提案手法では、既存の辞書項目を元に、学習データから抽出したパターンを用いて形態素解析を行う。実験では、複数の品詞タグ付きコーパスと辞書を用いて提案手法の評価を行い、最小コスト法や点推定に基づく形態素解析手法の既存実装と同程度の精度、 $1/2$ から $1/20$ 程度の消費メモリで、1,000,000文/秒を超える速度の形態素解析が可能なことを確認した。本手法の実装（C++で約1000行）は以下で公開する。

<http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/jagger>

NLPはChatGPTで終わるのか？

リツイート済み
 NLP2023 OKINAWA
@anlpmeeting ...

言語処理学会公式YouTubeチャンネルで
緊急パネル：ChatGPTで自然言語処理は終わるのか？
の動画を公開しました。

巨大言語モデルの出現でNLP研究はどう変わるので、GPT4リリース直前の
3月14日にNLP2023 OKINAWAの会場で行われた議論をぜひご覧ください。

youtube.com
NLP2023 緊急パネル：ChatGPTで自然言語処理は終わる...
概要：大規模言語モデルの発展によって自然言語処理
(NLP) の方法論は大きく様変わりした。中でもOpen AIか...

午後0:30 · 2023年4月14日 · 2.3万 件の表示

まとめ

- 現地での学会参加はめちゃくちゃ楽しい
 - 生で熱意のある発表が聞けるのは自分のモチベアップに
 - 発表の仕方, アイデアの面白さを学べる良い機会
- 知り合いを増やせた
 - 有志の懇親会に積極的に参加するとよい
 - 普段なら関わりにくいところに参加するとよい
- 次の学会参加に向けて
 - 自分から話しかけに行く
 - もっと知識を身につける
 - 面白い発表をする
 - 事前にどのセッションを見に行くか決めておく
 - 体力 is all you need.

おまけ：沖縄猫



おまけ：沖縄メシ

