LRE corpus アノテーションガイドライン

2016/09/07 Koji Matsuda

本文書は、場所参照表現タグ付きコーパスのアノテーションのためのガイドラインです。一部、作業手順書的な内容を含んでいます。本コーパスの目的は、ツイート中に含まれる実世界の場所を指し示すような表現に対して、

- テキスト中のどの表現が実際の場所を指し示しているか(Mention Detection : 言及抽出)
- その表現が指し示しているエンティティは、ガゼッタの中に存在するか、存在するとして、それはどれである可能性が高いか (Entity Resolution : エンティティ解決)

という2つのステップのアノテーションを行うことによって、ツイート中に含まれる実世界のエンティティを解決することを目指します。

原則

- 何らかのタグを付与すべきテキストスパン
  - 特定の場所を著者が想定している可能性が高い表現
- タグを付与しなくても良い表現
  - ハッシュタグ等
- 付与の単位は名詞句(固有名詞、普通名詞、連続する名詞のいずれか)。ただし、文節を超えることは無い
- 手がかりにして良い情報
  - ユーザーのプロフィール
  - 自分の地理的な知識
  - エンティティ解決においては、1メンション あたり1分程度を目安に時間をかけ, Google等で検索しながら付与するエンティティを決めても良い

Mention Detection : 言及抽出

このステップにおいては、 場所を指し示す表現を以下のクラスに分けてアノテーションします。この段階では、どのエンティティを指しているか、ではなく、どのようなエンティティを指しているか、という基準でアノテーション。

タグ	説明	例
facility	施設名	仙台駅、ハチ公前、東北大学
location	地名	仙台、渋谷、片平
rail	鉄道路線を表す表現	京浜東北線、仙石線、田園都市線
road	道路を表す表現	4号線、東北道、外環
generic	総称的な表現（今後の分析のため、「特定の」場所を指していない場所表現の一部にタグを付与しています）	英語の "a hospital" に訳されるような、不定の「病院」に対するメンション等
fiction	どこかの場所を指していることは確かであるが、現実世界の場所ではないもの	ゲームの中のエンティティ、小説の中のエンティティなど
other	場所を指している可能性はあるがそうでない可能性もあり、分析が難しいもの	「川崎のリーダー」という文脈における川崎（サッカーチームかもしれないし、川崎市かもしれない）

注： railカテゴリ, roadカテゴリは、論文中では facility タグと統合して分析を行っています。

タグ概要

タグの概要と、そこに含まれるような表現のうちの代表的なものを列挙します。

Location

- 場所を表す固有名詞(地名)
- 以下に当てはまる地名（領域を持つもの）
  - 国名、都市名
  - 自治体の名前 (宮城県、仙台市、太白区等)
  - 郵便の宛先の一部になりえる名前（一番町、川内）
  - ただし、ビルの名前は含まない（xトラストシティ）
  - 地方の名前（西東京、東三河、県北）

基本的には、 **人口** という属性を定義できそうなエンティティに対する参照表現を Location mention とみなす。

Facility

- 場所を表す固有名詞(施設名)
- 以下に当てはまる施設名
  - 鉄道駅、ビル、公園、ダムなどの施設、お店の名前
- 例
  - 宮の沢練習場、仙台市役所
  - お店の名前：イオン、フォーラス

rail

- 鉄道、路面電車などの路線
- 例: 田園都市線,東北新幹線,新幹線

road

- 道路(国道)、高速道路など、交差点の名前もこちらに含める
- 例: 国道一号, 東北自動車道, 東北道

other

- 一般には地名として扱わないような山の名前、川の名前
- 地名とも施設ともつかないが、著者が地理的なエンティティを想定していると思われる表現
  - 例: 実家、職場、バイト
  - これらは、後述のエンティティ付与の対象とします
- 施設部分名：レフトスタンド、3階、西口、○○ブース
  - いずれもother
- スポーツチーム等
  - とにかく打てない**広島_Other**
  - はな、**ロシア_Other**観戦いきますかね (ただし、実際にロシアに行こうとしていそうなら Location)
- 例)
  - **学校_Other**は楽しい
  - **会社_Other**変わった

アノテーション対象ではないもの

- 地名が含まれるが、全体では場所を指していない固有名詞
- 名詞句全体としては場所を表していない名詞
  - **東京フレンドパーク** 付与しない
  - **城下町のダンデライオン** 付与しない
  - **イタリア人** つけない
- 完全に(場所ではなく、)組織を念頭においているもの
  - **なか卯(付与しない)**と**すき屋(付与しない)**は同じ会社で運営していること

コーナーケース

- 架空のエンティティ
  - 現実世界に存在するとしたらどのようなものか、でつける。　XXX国: Location , YYY城: Facility
  - 現実世界にも存在しないもの：「チャットルーム」「ロビー」Other
- 海外の地名：できるかぎりアノテーションする
- 「○○(地名)××(施設名)」は繋げてアノテーション
  - **仙台二郎_Facility**
  - **仙台市役所_Facility**
- ただし、「○○ "の" ××」は分けてアノテーション
  - **仙台_Locationの二郎_Facility**
  - 「○○(地名)○○(地名)」は分けてアノテーション
  - **仙台市_Location"**青葉区_Location"**
- 接尾辞や接頭辞をもつことによって場所性を持つ複合名詞の場合は、それらを含めた範囲で付与する
  - 例) **パイト先_Other お風呂場_Other**
- **外国** Other
- **全国** Other,
- **東北地方** location

Entity Resolution : エンティティ解決

これらの表現に対して、エンティティを付与します。アノテーション補助システムにおいて、エンティティのリストは右ペインに表示されます。それぞれの表現に対して、ガゼッタに含まれる適切なエンティティに対して、チェックを行い、画面上部のボタンで適切と思われるエンティティを付与します。メンションを選択すると、それに応じた候補が自動的に右ペインに現れます。その中に適切な候補がある場合は、チェックボックスをクリックしてそのエンティティを選択してください。

- エンティティとして適切な可能性のある候補が複数見つかる場合は、それらの全てにチェックする。
  - しかしながら、「東京の**ドトール_FAC**」のように、可能性のある候補が多数（概ね10個以上）存在する場合は、それら全てをエンティティとして付与することは現実的ではないため、 **UNSP** タグを付与する。
- もし適切な候補が見つからない場合は、適切な候補が見つかるまで、1メンションにつき1分程度を目安に、クエリを繰り返し入力しながら、適切なエンティティを探す。
  - 今回は全文検索エンジンに Elasticsearchを用いていますので、そのクエリ言語が使えます。
  - もし、適切な候補が一つも見つけられない場合は、 Gazetteer には存在しないものと考えて **OOG** タグを付与する。

エンティティ解決の問題は、アノテーターの事前知識(居住地や地理的な知識に関するバックグラウンド等)によって結果が影響される可能性が高いが、今回のアノテーションでは問題としないことにする。

エンティティ付与の優先順位

地理的なエンティティには、地理的に包含関係が定義できる場合がある(例: 川崎駅は川崎区に含まれる、かつ、川崎区は川崎市に含まれる)。情報量という観点から、可能で妥当な解釈のうちで、最も Specificなエンティティに付与することにする。

たとえば、

> 川崎から京急川崎まで歩いてみる

というような文が与えられた場合、以下のようにアノテーションする。

> **川崎_FAC(JR川崎駅)**から**京急川崎_FAC(京急川崎駅)**まで歩いてみる

川崎市ではなく、川崎駅と付与。自然な解釈が成立するなかで最もスペシフィックなものに付与。

- 地名という解釈と、施設名という解釈の両方が可能であれば、施設名であると解釈してエンティティを付与する。
- 地名の解釈の中で、もしエンティティAの中に内包されるエンティティBがあり、その両方が候補になり得るのであれば、エンティティBを付与する。

優先順位	エンティティ	例
高	特定性の高い施設エンティティ	駅, ショッピングセンターなど
中	特定性の高い地名エンティティ	大字レベルの地名エンティティ
低	特定性の低い地名エンティティ	市区町村レベルの地名エンティティ
より低い	特定性の低い地名エンティティ	県レベル、地方レベルの地名

エンティティが一つに絞り切れない場合

- 基本的には全て付与
  - しかし、候補があまりにも多くなる(10個を超えるような)場合は **UNSP** を付与

その他

もし、エンティティ辞書に適切なエンティティが見つからない場合で、Wikipediaにその記事が見つかる場合は備考にその旨記載してください。