ECサイトにおける商品タイトルからの商品名抽出

Product Name Extraction from Product Entries on EC Pages

張　培楠（株式会社サイバーエージェント）

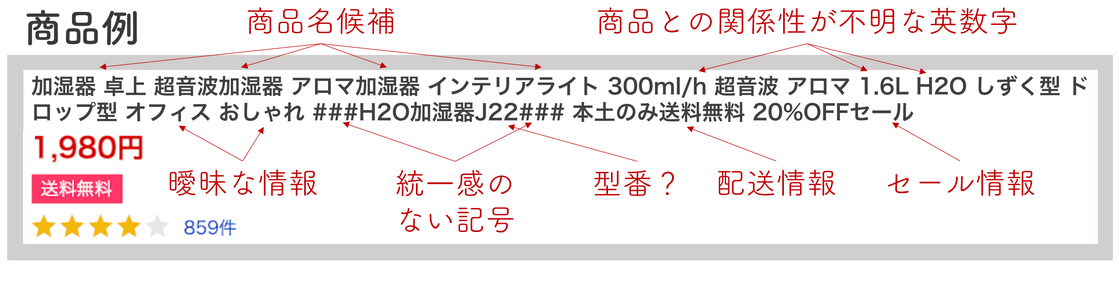
CyberAgent AI Lab

① Motivation

背景

EC が一般的になってきている中、出品されている商品のタイトルのほとんどは出品者が行っているため、過剰な情報付加により非常にわかりにくいものとなっている。

商品例



このような商品タイトルは購入者の視認性を損なっているだけでなく、運営者にとっても販売内容の把握を困難にする要因にもなっている。
従って雑多な商品タイトルからの商品名を抽出することは高い需要を持つ。

商品名

- 加湿器
- 超音波加湿器
- アロマ加湿器
- インテリアライト

こうした背景により、商品タイトルから商品名を抽出するタスクの考案とデータセットの作成、実験そして分析を行った。

② Data Creation

データ作成

18 の商品カテゴリから無作為に 15,000 件の商品タイトルを集め、分析を行った。右表は分析対象の商品タイトル一部。共通する特徴：

- タイトルの単語・フレーズが名詞・名詞句がほとんど
- タイトル全体で一貫した文法構造は持っていない
- 商品情報以外に商品とは直接関係しない配送情報や取引情報が数多く含まれる
- 表層が漢字、かな、英数字、記号が規則性なく混在

商品タイトル例

①	最大75%off！SALE開催中★[ロイネス]Roiness 犬用 いわし 150g /ウエット パウチ 国産 4571245858269 #w-148985
②	〔純正品〕 HP インクカートリッジ/トナーカートリッジ 〔C9363HJ HP134 3色カラー〕
③	ＴＲＵＳＣＯ　ナベ頭組込ネジ　クロメートＰ－４　サイズＭ６Ｘ１０　５０本入【メール便 送料300円選択可能です。】
④	（送料無料）（代引き不可）アーノルドパーマー　ベビー サンダル 13.5cm AP4111
⑤	当店１年保証 カシオCasio General Men's Watches Metal Fashion MTP-1183A-7ADF - WW

これらの商品タイトルから下のような制約を満たした 3,841 件の商品タイトルとその商品名のペアとなるコーパスを作成。

制約1: 連続した単語を選択
制約2: ひとつの単語のみを選択
制約3: より認識される最小単位の単語を選択

対象商品名

①	いわし
②	インクカートリッジ
③	ナベ頭組込ネジ
④	サンダル
⑤	Watch

コーパスの内訳

商品タイトル数	3,841
述べ文字数	220,900
異なり文字数	1,634
タイトル平均文字数	57.5

③ Methods

手法

Term Weighting (TF-IDF)

- 抽出対象の商品タイトルに含まれる単語に対してスコアリングしていくアプローチであり、ここでは TF-IDF を使用
- 「送料無料」や「代引不可」といった特定の商品に限定されない単語は多くの商品タイトルで出現するため、スコアは下がる
- 反対に特定の商品に関係する単語は限定的な商品タイトルにしか出現しないため、スコアは上がる
- TF-IDF は以下のように定式化

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d)\cdot \mathrm{idf}(t,D)$$
$$= \frac{\mathrm{freq}(t,d)}{\sum_{t_i\in d}\mathrm{freq}(t,d)}\cdot \log\left(\frac{|D|}{1+n_t}\right)$$

系列ラベリング

素性設計による手法 (CRF)

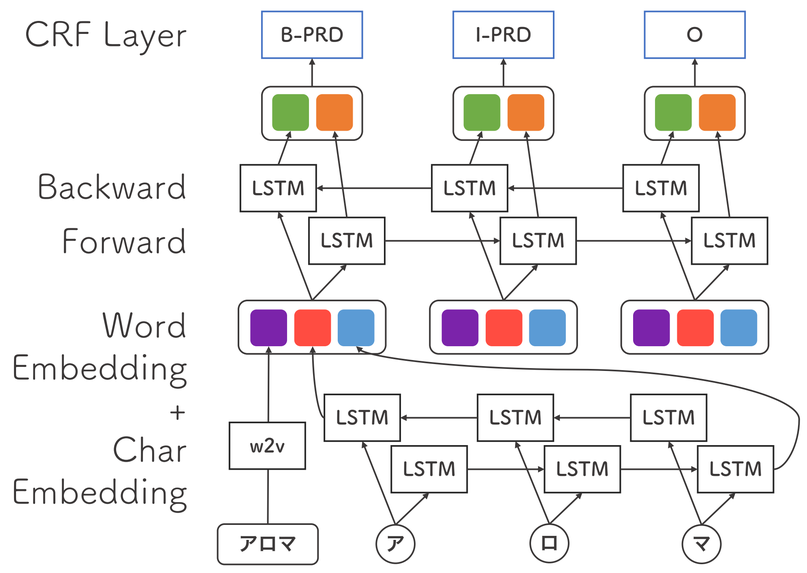
設計された素性

品詞素性	名詞や接続詞といった品詞情報
表層素性	漢字、かな、数字やアルファベットなど
位置素性	タイトルでの単語が存在している位置
辞書素性	形態素解析器の辞書に含まれているかどうか

上表のような素性を設計し、条件付き確率場（CRF）を用いてラベリングを行うアプローチ。CRF は以下のように入力系列 $\mathbf{X}$ に対して系列 $\boldsymbol{y}$ を予測する。

$$P(\boldsymbol{y}|\mathbf{X}) = \frac{1}{Z}\exp\sum_{t=1}^{n}\sum_{k=1}^{K}\lambda_k f_k(x, y_{\{t-i,\ldots,t\}}, t)$$
$$Z = \sum_{y\in\boldsymbol{y}}\exp\sum_{t=1}^{n}\sum_{k=1}^{K}\lambda_k f_k(x, y_{\{t-i,\ldots,t\}}, t)$$

ニューラルネットワークによる手法 (BiLSTM+CRF)



Lample らによって提案された単語・文字を BiLSTM を用いて得られた内部表現を CRF でラベル予測する手法（上図）。

④ Experiments & Discussion & Future Works

実験・考察・今後

BiLSTM+CRF による正答例および誤答例

1	正答例	Steve Madden (スティーブマッデン) メンズ Taslyn **ロー ファー** Grey
2	正答例	**防 ダニ ・ 抗菌 ・ 防臭 フランネルラグマット** / 絨毯 〔 ボリューム タイプ / 約 185 × 280 cm ピング 〕
3	両方	(業務 用 10 個 セット) 三甲 (サンコー) ベタ 目 **コンテナ ボックス** / **サンボックス** 11 - 2 (代 引 不可)
4	誤答例	**メンズ ブーツ** **チャッカブーツ** カジュアル Timberland Men ' s Groveton Leather Fabric Boot 正規 輸入 品
5	誤答例	日本 製 【 ■ **nano cafe** 】 ベビー **手 マグ** 54294 4522202542943
6	誤答例	HiKOKI (旧 日立工機) **フロア 用 タッカ** 電源 電圧 V **N 5004 MF** (送料 無料)(代 引き 発送 不可)
7	誤答例	ジョインテックス **応接** **センター テーブル** **KE - 1260 W**

凡例
- 予測箇所
- 正解箇所
- 正答箇所（予測と正解が被っている箇所）

実験結果

手法	適合率	再現率	F値
TF-IDF	5.7	11.1	7.0
CRF	20.3	11.4	14.5
BiLSTM+CRF	**25.4**	**21.2**	**23.0**

- ニューラルネットワークを用いた系列ラベリング手法の BiLSTM+CRF が F 値において他を大きく上回った
- 全体的に F 値が低く、難しいタスクであることが分かった
- 今回のデータでは、1 タイトルにつきひとつの正解ラベルしかつかないが、実際は複数の商品名候補が存在するので、複数ラベルを考慮したデータセットの作成が必要

⑤ References

参考文献

- Lafferty et al., Conditional random fields: Probabilistic models for segmenting and labeling sequence data. (ICML 2001)
- Lample et al., Neural architectures for named entity recognition. (NAACL-HLT 2016)

Paper

Poster

Coming Soon

About us