Towards Product Attributes Extraction in Indonesian e-Commerce Platform

Muhammad Rif'at, Rahmad Mahendra, Indra Budi, and
Haryo Akbarianto Wibowo

Faculty of Computer Science, Universitas Indonesia
Depok, 16424 West Java, Indonesia
muhammad.rifat@ui.ac.id, {rahmad.mahendra, indra}@cs.ui.ac.id

Abstract. Product attribute extraction is an important task in e-commerce domain. Extracting pairs of attribute and value from free-text product descriptions can be useful for many tasks, such as product matching, product categorization, faceted product search, and product recommendation. In this paper, we present a study of attribute extraction from Indonesian e-commerce product titles. We annotate 1,721 product titles with 16 attribute labels. We apply supervised learning technique using CRF algorithm. We propose combination of lexical, word embedding, and dictionary features to learn the attribute using joint extraction model. Our model achieves F1-measure 47.30% and 68.49% respectively for full match and partial match evaluation. Based on the experiment, we find that doing attribute extraction on more various number and diverse attributes simultaneously does not necessarily give worse result compared to extraction on less number of attributes.

Keywords: attributes extraction, e-commerce, product title, Named-Entity Recognition, Indonesian language

1 Introduction

In the recent years, e-commerce has gained rapid growth all over the world. Nowadays, online shopping is regarded as an integral part of our daily lives. E-commerce enables consumers to purchase a large number of products from a variety of categories, such as electronics, clothing, and foods.

An online marketplace features a very long-tail inventory and receives data from thousands of merchants about millions of products. Few sellers may provide rich and structured specification of product offers, while the vast majority only provides natural language description. Those product descriptions need to be extracted into pairs of attribute and value in order to gain useful information from them. Extracting attribute-value pairs from free-text data is an important building block as many tasks within the domain of e-commerce, including product matching, product categorization, faceted product search, and product recommendation, are able to take benefits from structured information about the products.

The products with specific functionalities are offered on the web with various brands or produced by different manufacturers. On the other hand, the same real-world products are sold by the number of e-shops, in which each retailer may provide heterogeneous product descriptions with different levels of detail. Given that many thousands of online shops sell millions of diverse products over the web, product matching has become of increasing importance. Extracting such attributes from the products can advance the matching process [3, 4, 12].

Another application that can utilize attribute-value pairs of product offer is content-based recommender systems. The recommender systems help the user to choose suitable item from many options in the product list; one of possible approach is content-based filtering. This approach tries to recommend items similar to those the user has liked / bought in the past. Having list of its attributes comes in handy to represent the content of a product [16].

Product attributes extraction problem can be formulated as Named-Entity Recognition (NER) task [13, 2]. Text fragments corresponding to particular attributes, e.g. brand and name, are considered as named-entity to be extracted. NER task in e-commerce domain faces several challenges. Product titles, which are typically short texts, do not necessarily follow grammatical structure. There are many abbreviations and typographical errors in writing the product descriptions.

Several studies have been conducted to tackle product attributes extraction problem using varied approaches: supervised and semi-supervised learning [1, 13, 8, 2], unsupervised learning [7, 15], and regular expression [11].

Previous work exploited data from large international e-commerce companies that uses English language. There has not been prominent work on attribute extraction from product offers in languages other than English. In fact, Asia Pacific is greatest B2C e-commerce market, based on survey by e-commerce foundation in 2015[1]. Of course, products being traded in the market are not only offered by global e-commerce companies, but also by local sellers. As textual description may be written in local language, attribute extraction from non English product offer is also useful task.

Indonesia e-commerce market has a lot of untapped potentials. By 2025, Indonesia is expected to dominate 52 percent of all e-commerce activities in Southeast Asia (Indonesias e-Commerce Landscape 2014: Insights into One of Asia Pacifics Fastest Growing Markets)[2]. This is due to the fact that Indonesia is one of the most populous countries in the world. Its geographical situation also makes e-commerce as an inevitable mode for distributing goods.

In this paper, we present sequence learning technique to extract multi-attributes from the product offers in Indonesian e-commerce platform. Some products that are available in Indonesian online market are imported goods. So, the products descriptions in our dataset are not only in Indonesian language, but also in for-

[1] https://www.ecommercewiki.org/wikis/www.ecommercewiki.org/images/5/56/Global_B2C_Ecommerce_Report_2016.pdf

[2] http://www.specommerce.com.s3.amazonaws.com/dl/wp/141215-white-paper-indonesia.pdf

eign language (mostly English). Our dataset also contains switch-code language product title.

In the remainder of this paper, we first discuss related work in the past. In section 3, we describe the dataset and annotation process. In section 4 and 5, we explain our proposed method and then report the result and analysis of our experiment. And in section 6, we conclude and discuss future work.

2 Related Work

Extracting attributes from product description has been done in some previous works. Mauge et. al. [7] conducted a work for structuring e-commerce items into descriptive properties (we call them attributes). They did unsupervised property discovery and supervised property synonim discovery using maximum entropy based clustering algorithm.

Ghani et.al. [1] extracted attribute-value pairs from apparel and sporting product textual description. They used bootstrapping method to get more training data from unlabeled data. The techniques applied for the attribute extraction are Expectation-Maximization (EM) and Naive Bayes.

Putthividhya and Hu [13] extracted four types of attributes (i.e. `brand`, `garment type`, `size`, and `color`) from e-Bay clothing and shoes product categories. They also compared several methods to extract attributes, such as Hidden Markov Model (HMM), Maximum Entropy, Support Vector Machine (SVM), and Conditional Random Field (CRF). The result showed that CRF give better result. Putthividhya and Hu also used bootstrapping, however, not to add training data, but to expand the seed list of attribute values.

Radhakrishnan et.al. [14] parsed Amazon product titles from the camera and photo categories to obtain several semantic tags: `brand name`, `product name`, and `version`. They used CRF to tackle the problem. More [10] worked on `brand` recognition from Walmart product titles. More‘s system combined sequence labeling algorithms, such as CRF and Structured Perceptron, with a normalization scheme.

Joshi et.al. [2] investigated use of distributed word representation as features for NER task in e-commerce domain. They found that extending basic features (i.e. lexical and orthography) with word embedding that is pre-trained on combination of in and out-of-domain corpus, using CRF, delivers best result. Joshi et.al. conducted the experiments on e-Bay products from 5 categories, i.e. cellphones, cellphone accessories, men‘s shoes, watches, and women‘s clothing.

Kozareva et.al. [5] studied the performance of multiple structured prediction algorithms to automatically recognize `product`, `brand`, `model`, and `product family` semantic types from the shopping queries.

Melli [8] proposed annotation structure to apply to the product titles from diverse set of industry types. Each product title is chunked into semantically classified sub-segments, that are product identifying term, product feature term, product category term, product brand term, product line term, merchant term, offering feature term, and functional term. To parse the product title, a supervised CRF model was used.

Our work, similar to Melli [8], is conducted on the dataset of products from a broad range of categories. However, instead of abstract concept term, our gold-standard data is labeled with familiar high-level semantic categories, e.g. `brand`, `name`, and `size`.

3 Data and Annotation

There are different forms of textual information associated with an e-commerce item. Product specification and long text description are often optional, vary greatly from seller to seller and among marketplaces. While, the product title is mandatory information for most marketplaces. It is indexed by the search engine and searched against by users of the website. Snippets shown in search result pages are generated from the product titles. In this work, our focus is only extracting attributes from the product titles.

We collected the data from 3 different Indonesian e-commerce websites, namely Elevenia[3], Bukalapak[4], and Lazada[5]. We obtain 91,395 Elevenia, 26,728 Bukalapak, and 53,417 Lazada product titles. Due to resources limitation, only a small portion of data is annotated and used for the experiment. The sample of the data is selected by stratified sampling technique based on product category.

The products categorization varies among several e-commerce platforms. Each marketplace has its own product taxonomy. For example, Elevenia classified the products in its website into 8 categories. While, each product in Bukalapak e-shop is categorized into nearly 20 fine-grained classes. To deal with differentiation of categories among platforms, we applied rule-based mapping approach. After the mapping process, we consider only products from top-15 frequent categories, i.e. Fashion, Electronics, Handphone & Tablets, Computer & Laptop, Foods & Drinks, Health & Beauty, Baby & Children, Home Application, Office Supplies, Hobby, Jewelry & Watch, Automotive, Industry Tools & Appliances, Media, and Gadget Accessories. Our sample data consists of 1,721 product titles.

We define 16 kinds of attributes in our annotation scheme, as followings:

1. `Brand` attribute. A brand is a trademark or distinctive name identifying the product. The brand makes a product distinguishable from a clutter of products in the marketplace. It can be a company brand, an endorsed brand, or an individual product brand.
2. `Name` attribute. The name is given to the variant of the product that is released in a certain period and/or area. The product name attributes in our annotated data can be a series number or the brand naming extension.

[3] `http://www.elevenia.co.id`

[4] `http://www.bukalapak.com`

[5] Lazada is actually an international e-commerce company. Our research uses only the data from Indonesian local site `http://www.lazada.co.id`

3. `Type` attribute. It specifies the functionalities of the product, e.g. *television*, *kitchen knife*, and *lotion whitening.* The product type attribute may be associated with the fine-grained classification of the product category in particular e-shop.
4. `Color` attribute. The attribute is any term that describes the product‘s color. The terms are not limited only to the basic colors (e.g. *green*, *blue*), but are also list of exhaustive color (e.g. *navy*, *copper gold*).
5. `Size` attribute. The attribute describes the standardized size of the product. It can be numeric or nominal data (e.g. *small*, *medium*)
6. `Gender` attribute. Some products are targeted to be sex-specific. Gender-based information may be mentioned in the product title, as in the following example, *Polo shirt for men.*
7. `Age` attribute. Some products are designed for specific age group. Mention of the age information in the product title can be numeric, interval, or nominal (e.g. *toddler*, *adult*).
8. `Material` attribute. The attribute describes the main fabric or material that the product is made of (e.g. *plastic*, *cotton*).
9. `Pattern` attribute. Several products may have the pattern feature in the design. For example, *striped shirt.*
10. `Theme` attribute. The attribute describes the graphic printed on the product. A theme usually comprises a set of texts, shapes, and colors. It can be logo, painting, photo, or title of the popular entity (such as film, public figures, sport team). For example, *thermos bottle hello kitty.*
11. `Shape` attribute. This attribute describes the physical form of a product. For example, *cetakan kue ikan koi* (translation: fish-shaped cookies mold).
12. `Mass` attribute. It is the unit weight of a product.
13. `Dimension` attribute. It can be the length, width, or volume of the product.
14. `Quantity` attribute. It is the number of pieces of individual item in a product package.
15. `Flavor` attribute. This attribute is typically present in foods, also health and beauty products It defines the flavor of product that varies by taste.
16. `Misc` attribute. Some products, e.g. electronics or computers, have a long list of specification. Typically, the specification is specific towards product category. For example, camera has `flash model`, `shutter speed`, and `file format`; refrigerator has `door type` and `frozing system`. In our annotation, the variation of product specification is considered as `Misc` attribute.

An attribute can be represented as single or multi-tokens within product title. On the other hand, not all tokens in the product title correspond to any of 16 pre-defined attributes.

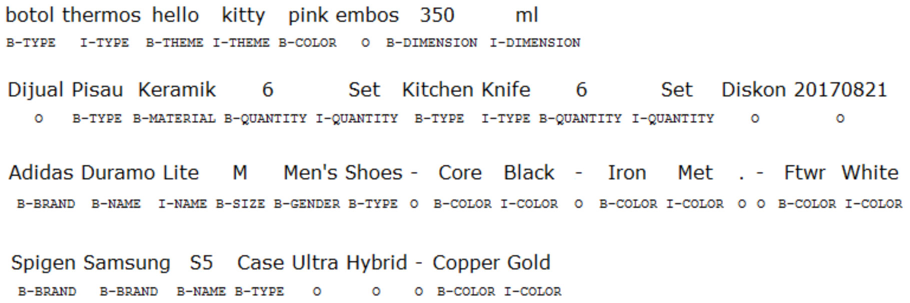
The first step of annotation process is tokenizing the product title into a sequence of tokens. The whitespace character is used as token boundary in general. Each single non-alphanumeric character is considered as an individual token, except for these following cases:

- a comma or a full stop occurring in between the sequence of numbers,

- a hyphen symbol occurring in between of the sequence of alphabet words or the sequence of numbers
- an apostrophe preceding / in between of the sequence of alphabets, or following the sequence of numbers

The sequence of tokens in the product title are labeled with BIO encoding. For a sequence of tokens corresponding to a particular attribute X, the label `B-X` is assigned to the first token in that sequence, while any token other than the first is labeled with `I-X`. The label `O` indicates that the token is not part of any attribute in the corresponding title. In our case, with 16 attributes, each token must be labeled with one of 33 possible labels: {`O`, `B-Brand`, `I-Brand`, `B-Name`, `I-Name`, ..., `B-Misc`, `I-Misc`}.

Figure 1 shows the example of the labeled product title.



botol thermos hello kitty pink embos 350 ml
B-TYPE I-TYPE B-THEME I-THEME B-COLOR O B-DIMENSION I-DIMENSION

Dijual Pisau Keramik 6 Set Kitchen Knife 6 Set Diskon 20170821
O B-TYPE B-MATERIAL B-QUANTITY I-QUANTITY B-TYPE I-TYPE B-QUANTITY I-QUANTITY O O

Adidas Duramo Lite M Men's Shoes - Core Black - Iron Met . - Ftwr White
B-BRAND B-NAME I-NAME B-SIZE B-GENDER B-TYPE O B-COLOR I-COLOR O B-COLOR I-COLOR O O B-COLOR I-COLOR

Spigen Samsung S5 Case Ultra Hybrid - Copper Gold
B-BRAND B-BRAND B-NAME B-TYPE O O O B-COLOR I-COLOR

Fig. 1. Example of Annotated Product Titles

The most frequent attribute label in our annotation is `Type`. Attributes `Type`, `Brand`, and `Name` occur respectively 2,003, 1,202, and 887 times. While, attributes `Shape` and `Flavor` are the most infrequent ones. Distribution of number of attribute annotated in our data is presented in Figure 2.

4 Method

We formulate product attributes extraction as a sequence labeling problem, in which each token w_i is associated with a hidden label $y_i \in$ set of attributes. Formally, given a product title containing N tokens $W = (w_1, w_2, ..., w_N)$, we want to find the best sequence of labels $Y = (y_1, y_2, ..., y_N)$, in which each label is determined using probability $P(y_i|w_{i-l}, ..., w_{i+l}, y_{i-l}, ..., y_{i+l})$; and l is a tiny integer.

We approach the problem with supervised learning using Conditional Random Field (CRF) [6]. Features used for the experiment are combination of lexical, word embedding, and dictionary features. Lexical features include bag-of-words,

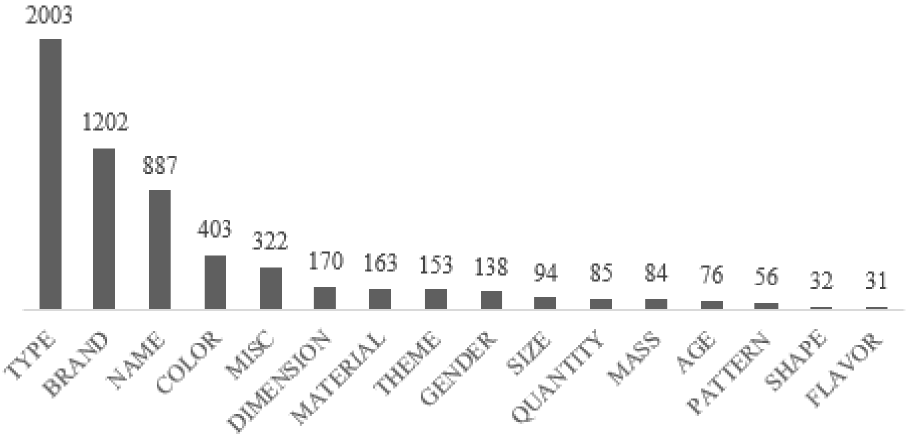


Fig. 2. Frequency Distribution of Attributes in Annotated Product Titles

orthography, and word position, while dictionary features include list of attribute values and language code. Information about lexical and dictionary features are listed in Table 1.

Table 1. Lexical and Dictionary Features.

Feature	**Description**
Bag-of-Words (`bow`)	current token and tokens in window of two (w_{-2}, w_{-1}, w_0, w_1, w_2)
Orthographic (`orth`)	Orthographic feature of current token and tokens in window of 1, e.g.: `CamelCase`, `Containdigit`, `ALLCAPS`, etc.
Word Position (`pos`)	Relative position of token in product description
Attribute Dictionary (`attrDict`)	Occurrence of token in dictionary of feature
Language Code (`lang`)	Possible language of current and surrounding tokens, checked based on language code dictionary (e.g.: the word "*knife*" exists in English dictionary, but word "*pisau*" does not)

To obtain word embedding (`we`) feature, we use skip-gram model from word2vec [9] pre-trained by in-domain corpus, i.e. corpus of e-commerce product titles itself. Distributed word representation features that are taken into consideration are current token and tokens in window of 2 (w_{-2}, w_{-1}, w_0, w_1, and w_2).

5 Experiment and Result

5.1 Evaluation Setting

We conduct the experiment with 10-fold cross validation, each of which adopts a 90-10 split. 90% of the data is used for the training, while the other 10% is for the testing.

The result of attribute extraction are evaluated in entity-level with full-match and partial match methods. To illustrate, suppose that the expected label consists of two tokens or more and the predicted contains only one token of them, partial match still counts it as a true positive. While, full match considers the prediction for multi-token attributes is true if all those tokens are completely predicted and no other token is included in the predicted label. As for evaluation, we measure precision, recall and F1 metric.

5.2 Feature Ablation

Feature ablation aims to determine the contribution of the features toward performance of extraction model. To measure contribution of each feature, model that applies all proposed features is firstly evaluated. Then, a feature is ablated one by one from the model. As there are 6 feature groups, we create 6 new different models, each of which uses remaining 5 features after ablation. Each of 6 models are evaluated and compared to model with all features.

Feature ablation study is reported in Table 2.

Table 2. Feature Ablation Study

Features	Full Match (%)			Partial Match (%)		
	precision	recall	F1-Measure	precision	recall	F1-Measure
all	**52.66**	**42.93**	**47.30**	76.26	**62.16**	**68.49**
all - `wordEmb`	51.91	42.50	46.73	75.66	61.95	68.12
all - `bow`	49.91	39.24	43.94	74.21	58.34	65.33
all - `orth`	51.61	38.62	44.18	**78.25**	58.56	66.99
all - `pos`	51.74	42.16	46.46	75.98	61.91	68.23
all - `attrDict`	51.13	40.88	45.43	75.77	60.57	67.32
all - `lang`	51.57	41.83	46.20	75.94	61.60	68.02

Based on the ablation study, it can be seen that all proposed features contribute positively to increase model performance. When one of them is removed in the experiment, F1-Measure decrease. The most influential feature in our model is the bag-of words.

5.3 Analysis of Model Performance

If the evaluation metric is broken down more detail into the attribute class, we find that number of attributes in the data has correlation with the model performance in predicting the attributes. Frequent attributes (e.g. `brand`, `color`) are relatively able to be predicted more accurately than rare attributes (e.g. `shape`, `flavor`). Evaluation of each attribute is reported in Table 3.

Table 3. Evaluation of Attributes Extraction from Product Title

Attribute	Full Match (%)			Partial Match (%)		
	precision	recall	F1-measure	precision	recall	F1-measure
`brand`	65.71	61.27	63.28	74.44	69.42	71.69
`name`	34.67	31.54	32.93	65.25	59.47	62.05
`type`	41.91	34.39	37.77	80.93	66.41	72.93
`color`	75.37	65.19	69.84	87.61	75.71	81.15
`size`	58.12	31.87	40.54	69.40	37.22	47.68
`gender`	91.29	88.22	89.59	91.29	88.22	89.59
`age`	48.98	33.50	38.06	56.16	38.31	43.68
`material`	55.27	26.62	35.22	78.36	36.74	49.17
`pattern`	56.33	26.65	33.28	70.33	33.01	41.64
`theme`	38.70	13.69	19.56	58.11	21.74	30.54
`mass`	93.18	75.06	81.86	95.52	76.68	83.72
`dimension`	62.79	55.85	58.43	65.67	58.41	61.13
`quantity`	63.39	40.25	48.67	80.31	52.67	62.77
`flavor`	20.00	15.00	16.67	20.00	15.00	16.67
`shape`	5.00	1.67	2.50	5.00	1.67	2.50
`misc`	64.55	41.85	50.36	75.95	48.85	58.98

`Gender` is the attributes with highest F1-measure (89.59%). Token representing `gender` attribute does not vary much. In addition, representation of `gender` is usually only single token, so it is reasonable if the prediction result is good and there is no difference between full match with partial match.

`Color` is the third best extracted attributes in prediction model. The fact that the `color` attribute extraction has a convincing score can be caused by at least two reasons. First, most entities of color attributes consist only of at most two tokens, and in general only one. Second, many colors appear on products title are popular, so they are repeated many times in the data. As most product title contain `color` attributes and the variation of those attribute values are not too broad, the bag-of-words and word embedding features are able to predict them quite well.

The attributes containing number format token, such as `mass`, `quantity`, and `dimension`, perform well enough. Such attributes have characterized orthographic pattern. The attribute is usually composed of a number followed by 2 or 3 letters denoting the unit. For example, *300gr* (`mass`), *3pcs* (`quantity`), and *180x200x20cm* (`dimension`).

Extraction of several attributes, i.e. `type`, `name`, `theme`, and `pattern`, obtains the evaluation score with wide gap between full match and partial match. The length of those attributes varies from 1 to more than 3 tokens. Such attributes consists of open-class words (most of them are OOV), so it is very challenging for the model to recognize them perfectly from a product title.

Our model is a joint extraction model that recognize multi-attributes simultaneously in just one process. While, most previous works on attribute extraction tasks focus only on limited number of attributes. We want to see the changes of model performance if we reduce number of attributes extracted from product title. We conduct the experiment to compare between a joint model with exhaustive list of attributes, a joint model with limited number of attributes, and the binary model. Model performance comparison is evaluated on three main attributes, i.e. `brand`, `name`, and `type`.

Another joint model is implemented with only considering those three attribute labels. Other attributes are treated as `O` label in this model. We also implement three binary models, that works separately to learn each attribute. Model comparison is reported in Table 4.

Table 4. Comparison of Attribute Extraction Model

Attribute	Model	Full Match (%)			Partial Match (%)		
		precision	recall	F1-measure	precision	recall	F1-measure
`brand`	joint-16attr	65.71	**61.27**	63.28	74.44	**69.42**	71.69
	joint-3attr	68.15	60.16	63.79	77.23	68.16	72.28
	binary	**75.12**	55.20	63.48	**82.19**	60.53	69.55
`name`	joint-16attr	34.67	**31.54**	32.93	65.25	**59.47**	62.05
	joint-3attr	38.64	29.10	33.06	71.27	53.58	60.93
	binary	**47.99**	23.90	31.78	**77.50**	38.47	51.22
`type`	joint-16attr	41.91	**34.39**	37.77	80.93	**66.41**	72.93
	joint-3attr	44.27	33.69	38.24	83.29	63.42	71.97
	binary	**47.03**	31.30	37.56	**86.20**	57.40	68.86

It is intuitive that binary model has higher precision compared to the joint model. However, extracting multi-attributes with joint model can obtain better recall. In the end, we believe that extracting abundant number of attributes simultaneously still achieve satisfying performance.

6 Conclusion

We presented attribute extraction task from product title in Indonesian e-commerce data. We annotated 1,721 sample data from 3 different Indonesian e-commerce platforms. Our data, that consists of 15 various product categories, is annotated with 16 attribute labels. We approach product extraction problem as sequence labeling task. We apply supervised learning using CRF algorithm in this study. Six feature groups are proposed, i.e. bag-of-words, word position, token orthography, word embedding, language code, and attribute dictionary features. Our model extract all attributes jointly in a single learning process. The experiment shows that performance of attribute extraction does not much affected by number of attributes to extract. Doing multi-attributes extraction from the product title using the joint model is worth the effort to obtain more structured information about the product.

There are a couple of interesting directions that may be considered for future research. First, even though our data consists of mixed of Indonesian and English language, analysis of model related to switch-code data have not covered in this study. As e-commerce products are traded in international market, model adaptability in cross-lingual and multi-lingual setting is useful. Second, as we approach the problem as named-entity recognition task, each token can be annotated exactly with one label. However, we find several cases in which a token can correspond into more than one label (e.g. *girl* is not only interpreted as `gender`, but also implicitly embeds information about `age` attribute). Third, this study has not completely extracted ready-to-use product attribute. Further step to identify semantically equivalent values amongst extracted attributes is needed. Last but not least, considering current performance, the learning model is widely open for improvement.

Acknowledgement

The authors gratefully acknowledge the support of the PITTA UI Grant Contract No. 406/UN2.R3.1/HKP.05.00/2017. The first author was also partially funded by Bukalapak

References

1. Ghani, R., Probst, K., Liu, Y., Krema, M., Fano, A.: Text mining for product attribute extraction. SIGKDD Explor. Newsl. 8(1), 41–48 (Jun 2006)
2. Joshi, M., Hart, E., Vogel, M., Ruvini, J.D.: Distributed word representations improve ner for e-commerce. In: VS@ HLT-NAACL. pp. 160–167 (2015)
3. Kannan, A., Givoni, I.E., Agrawal, R., Fuxman, A.: Matching unstructured product offers to structured product specifications. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 404–412. KDD '11, ACM, New York, NY, USA (2011)
4. Köpcke, H., Thor, A., Thomas, S., Rahm, E.: Tailoring entity resolution for matching product offers. In: Proceedings of the 15th International Conference on Extending Database Technology. pp. 545–550. ACM (2012)
5. Kozareva, Z., Li, Q., Zhai, K., Guo, W.: Recognizing salient entities in shopping queries. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers (2016)
6. Lafferty, J.D., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. pp. 282–289. ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001)

7. Mauge, K., Rohanimanesh, K., Ruvini, J.D.: Structuring e-commerce inventory. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. pp. 805–814. Association for Computational Linguistics (2012)
8. Melli, G.: Shallow semantic parsing of product offering titles (for better automatic hyperlink insertion). In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1670–1678. ACM (2014)
9. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
10. More, A.: Attribute extraction from product titles in ecommerce. arXiv preprint arXiv:1608.04670 (2016)
11. Petrovski, P., Bryl, V., Bizer, C.: Learning regular expressions for the extraction of product attributes from e-commerce microdata. In: Proceedings of the Second International Conference on Linked Data for Information Extraction-Volume 1267. pp. 45–54. CEUR-WS. org (2014)
12. Petrovski, P., Primpeli, A., Meusel, R., Bizer, C.: The wdc gold standards for product feature extraction and product matching. In: International Conference on Electronic Commerce and Web Technologies. pp. 73–86. Springer (2016)
13. Putthividhya, D.P., Hu, J.: Bootstrapped named entity recognition for product attribute extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1557–1567. Association for Computational Linguistics (2011)
14. Radhakrishnan, P., Gupta, M., Varma, V.: Modeling the evolution of product entities. In: Proceedings of the 37th International ACM SIGIR Conference on Research; Development in Information Retrieval. pp. 923–926. SIGIR '14, ACM, New York, NY, USA (2014)
15. Shinzato, K., Sekine, S.: Unsupervised extraction of attributes and their values from product description. In: Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013. pp. 1339–1347 (2013)
16. Xu, B., Zhang, M., Pan, Z., Yang, H.: Content-based recommendation in e-commerce. In: Computational Science and Its Applications – ICCSA 2005. pp. 946–955. Springer Berlin Heidelberg, Berlin, Heidelberg (2005)