Accurate Product Name Recognition from User Generated Content

Sen Wu*
Department of Computer Science
Tsinghua University
Beijing, China
ronaldosen@gmail.com

Zhanpeng Fang*
Department of Computer Science
Tsinghua University
Beijing, China
fzp1990@gmail.com

Jie Tang†
Department of Computer Science
Tsinghua University
Beijing, China
jietang@tsinghua.edu.cn

This paper presents the solution of the team "ISSSID" for the Consumer PRODucts Contest #1(CPROD1) of ICDM 2012[1]. The contest provides a dataset including hundreds of thousands of text items, a product catalog with over fifteen million products, and hundreds of manually annotated product mentions. The goal of the competition is to automatically recognize product mentions in the textual content and disambiguate which product(s) in the product catalog are referenced by the mentions. We propose a hybrid approach which combines the results obtained by several separately trained recognition models. Specifically, the approach uses a standard matching model, a rule template model, and a conditional random field model, and finally combines the results using a blending model. The proposed approach achieves the best performance in the contest.

Nature Language Processing, Named Entity Recognition, CPROD1

I. Introduction

Internet plays an important role in people's daily life. A significant proportion of web usage is to acquire information, discussions, researches, and purchase of consumer products. Indeed, people nowadays are strongly influenced by social users' opinions. For example, users usually want to first refer to the others' comments, before purchasing a product. Thus it would be very useful if we could design a service which is able to extract product related information and align them to various products.

However, one challenge here is the disambiguation problem of product names on the Web. A product may have multiple different names, e.g., abbreviated name, full name, while different products may have the same name, e.g., "Apple". More general, given a collection of documents, whether and where a product is referred to? The problem is related to a research problem called Named Entity Recognition (NER) [1], in which the goal is to find the corresponding "correct" information from large-scale data for a user given keyword (e.g., a product name).

ICDM-2012 CPROD1 Contest aims to solve such a problem. Specifically, its goal is to determine the state-of-the-art methods to (1) automatically recognize product mentions from textual content and to (2) disambiguate which product(s) are being referenced. Figure 1 shows an example to demonstrate the task in the contest. In general, there are two subtasks. The first subtask is to identify product mentions from the textual content and the second subtask is to align the identified mentions to product entries in the product catalog. Comparing with the traditional NER, the problem has several unique challenges.

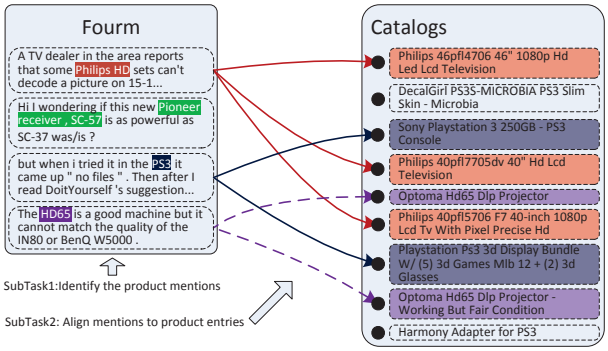


Figure 1. Product name recognition from user generated content. In general, there are two subtasks. The first subtask is to identify product mentions from the textual content and the second subtask is to align the identified mentions to product entries in the product catalog.

1) **Heterogeneous data.** Besides the training data, the dataset also contains some heterogeneous data including product name, product category (consumer electronics (CE) or automotive (AU)) and the price of each product. How to utilize the heterogeneous information to enhance the performance of identification is a critical issue.
2) **Semantic behaviors.** The contest dataset were collected from people's conversations in forums. The text contain much noise. How to extract the semantic information from the noisy data is a challenging problem.
3) **Identify products.** In the contest, each participation needs to not only recognize all mentions of the consumer products in a large user generated collections of web-content, but also align each mention to the right product in the catalog of products. Both the recognition and the alignment requires high accuracy.

*The author ordering does not indicate differences in contributions.
†Jie Tang is the advisor of the team.
[1]http://icdm2012.ua.ac.be/content/contest

To solve the above challenges, we propose three models: *Standard Match model*, *Rule Templates model* and *Conditional Random Field model*. We use *Standard Match model* to identify the products presented in the training dataset. The *Rule Templates model* leverages the products naming rules and several semantic information. We also propose a *Conditional Random Field model* to train the potential pattern which can not be provided by simple statistic analysis. As these models leverage different information, we propose a hybrid approach which combines the results of different models. The approach achieves an F-measure of 0.22041 on the private leaderboard, which takes the first place in the Contest of ICDM-2012. Our technical contributions are as follows:

- The proposed approach leverages the heterogeneity of the input data.
- We consider human semantic behaviors to enhance the performance.
- We design a very efficient matching method to identify the products from a large catalog of products.

The rest of this paper is organized as follows: Section II presents the proposed general framework; Section III Section IV gives the experimental results; Section V concludes the paper.

II. Approach Framework

The proposed solution follows three steps: Modeling, Blending and Recognition, as shown in Figure 2.

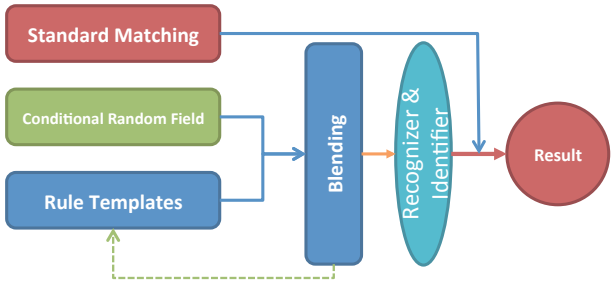


Figure 2. Approach Framework.

First, three models are designed and implemented separately to find product mentions. This step is called Modeling. Second, we incorporate the individual models via a blending method and filter the blended results, thus this step is called Blending. Finally, recognition method is used to identify the products referred by each mention we find.

III. The Proposed Models

In this section, we describe the three models used in our approach, i.e., Standard Matching model, Rule Templates model, and Conditional Random Field model.

Table I
General List

Categories	Examples
GD	English words in General dictionary
SW	Stop words (e.g., his, her)
CN	Capitalized nouns (e.g., January, Monday)
CA	Common abbreviations (e.g., mins, kg)

Table II
Features defined in the CRF model. The first twelve features are basic features for all CRF models. * represents features used by the additional CRF model.

Features	Examples
TOKEN	The current token
FC	If the first character is upper-case/lower-case
CHARCNT	The number of characters
UCCNT	The number of upper-case characters
NUMCNT	The number of numeric characters
LCCNT	The number of lower-case characters
DSHCNT	The number of dash-characters
SLSHCNT	The number of slash-characters
PERIODCNT	The number of period-characters
GRWRDCNT	The number of of matching grammatical words
BRNDWRDCNT	The number of matching English common words
ENWRDCNT	The number of matching brand words
P_TOKEN*	The previous token
P_PREP*	If the previous token is a preposition
PF*	Pattern features introduced by M. Collins et al. [5]

A. Standard Matching Model

The easiest way for identifying the names of products is to use the annotated information in the training data. Based on this idea, we design the Standard Matching model. Specifically, we simply extract those terms/symbols that are annotated as products in the train data, and then find their occurrences in the test data. If a document contain the corresponding terms/symbols of a product, we say there is a mention of the product.

B. Rule Templates Model

In the Rule Templates model, candidates are recognized by a set of rule templates. Each rule is used to identify relevant entities. For example, the Nokia corporation has a series of cell phones named as 'N#' where '#' represents a number, for example 'N97'. The rules are defined as follows:

Special Words Product naming is a critical and creative process. Many factors will be considered, aiming to batch a product's shorthand. Basically, linguistic information is very important. For example, many products' names use the combination of specific characters such as 'iPhone', 'ThinkPad'. Based on the observation, we consider the only-one-gram non-standard words which appear no more than 20 times as the "special words". In total 4 million "special words" in the data set.

Semantic Patterns Base on the analysis of people's habits, we summarize three important patterns:

1) A product name always follows a pronoun, preposition

Table III
PERFORMANCE OF DIFFERENT METHODS.

No.	Models	Public Leaderboard	Private Leaderboard
1	Standard Matching	0.14557	0.09005
2	Rule Templates	0.15844	0.11365
3	CRF1	0.16328	0.15775
4	CRF2	0.12168	0.14390
5	3 + 4	0.17375	0.17465
6	1 + 2	0.26525	0.17909
7	6 + 3	0.30656	0.20526
8	7 + 4	0.30379	0.22041

or quantifier, such as 'my mac', 'the Xbox', 'one GTR' and so on.

2) When a sentence mentions some products and contains preposition 'for' in it, the words before 'for' has a higher probability to be the name of a product than those after 'for'. For example 'BlackBerry Curve 8900' is not a mentioned product while 'Seidio Innocase 360' is a mentioned product in sentence 'Seidio Innocase 360 for BlackBerry Curve 8900'.
3) The words following 'by' usually represent a user name or an organization rather than a product name. For example we should ignore 'jbooker82' in sentence 'Posted by jbooker82' .

General list Table I summarizes several categories of words that are rarely used for naming products names.

Based on the above rule templates, we propose a rule template model to identify the product names. The model is a cascade model. Each rule template is considered as a classifier and the entire model is based on the concatenation of all the classifiers. All the final words classified correctly by the cascade rule template model are the symbols of some products.

C. Conditional Random Field Model

The Conditional Random Field (CRF) model follows the version introduced by Andrew McCallum [2]. The CRF model allows both discriminative training and the bi-directional flow of probabilistic information across the sequence. It is used to represent the probability of a hidden state sequence by encoding known relationships between observations and construct consistent interpretations. It is widely used for labeling and parsing the sequence data in nature language processing. Finkel et al. [3] proposed a CRF sequence model named Stanford Named Entity Recognizer[2]. Faruqui et al [4] presented the best systems for German NER based on CRF and etc.

Our approach is similar to baseline 2[3] provided by the contest. It trains a sequence tagging model that classifies each token as one of the categories "I", "O", and "B". The letter "B" indicates that the token is the beginning of a product mention; the letter "I" indicates that the token is inside a product mention; and the letter "O" indicates that the token is outside a product mention. In our final submission, we use two different CRF models to identify the product mentions. Table II lists all the features used in the CRF models. Note that to support GRWRDCNT, BRNDWRDCNT and ENWRDCNT, it needs a organizer-provided dictionary file (dictionary.dat) with 86,024 entries of grammatical words, common English words, and brand names from the consumer electronics and automotive domain.

[2]http://nlp.stanford.edu/software/CRF-NER.shtml

[3]http://www.kaggle.com/c/cprod1/forums/t/2287/crf-based-baseline2-published

D. Blending Method

All the models proposed so far focus on different aspects of the problem, thus the mention symbol candidates blending may be helpful. The blending model we proposed is to filter duplicated candidates. Because the candidates mentioned by CRF models may be not recognized by the rule templates, we filter the candidates by rule templates method.

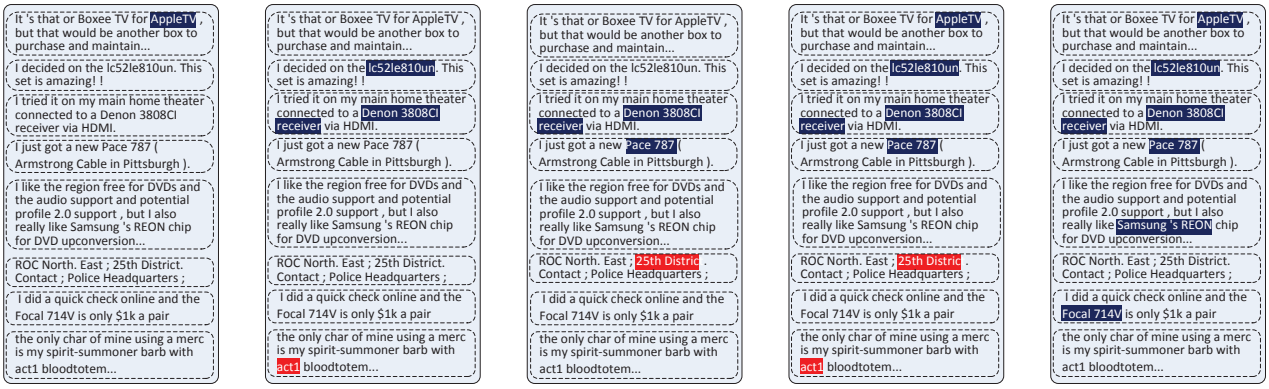
E. Retrieve and Recognize Products Names

After we get the mention symbols, we use an interactive mechanism to recognize the whole products name and retrieve the product items. For each mention symbol candidate, we first construct a name characters set using the products data which contains the symbol. Then we expand the symbol on both sides if the neighbor characters are in the set. After we get the whole product name, we try to identify which product items the name belongs to. The identification method is similar to the method before. We simply select the product items with name containing the mentioned product name. Intuitively, every product name should only belong to one category, either consumer electronics (CE) or automotive (AU). Our final work is to determine which category the product belongs to. We tried several methods, for example, voting method, semantic method (consider the whole sentence) and weighting method. Eventually, we use the voting method that simply choose the category which gets the most votes from the product items as our approach.

IV. EXPERIMENT

Validation Set

As the contest has a limit of only two entries per day submission limits, it is necessary to make a reasonable



(a) Standard Matching (b) Rule Templates (c) CRF (d) Blend (e) Ground Truth

Figure 3. Comparison of different methods for product name recognition. The text highlighted with dark blue indicates a correct recognition by a method and text highlighted with red color indicates a wrong recognition.

validation set for the competition, which is also useful to avoid overfitting problem. As for the Rule Templates model, it does not need a training data set, we then use the whole training set as the validation set. We found that performance of our approach on the validation set is very consistent with that in the public leaderboard. Therefore, this gives us high confidence to use the validation set to evaluate the performance of individual models and the blending method.

Results and Discussions Table III shows the results of different methods on the data set. From the table, we can see the highly consistent relationship between public leaderboard and private leaderboard. Our method significantly improves the performance of recognition. While the performance of each individual model is limited, the combination can significantly improves the performance. For example, by combining the standard model and the rule template, we could achieve a significant performance improvement (11-12%). By further incorporating the power of of the CRF model (i.e., the "6+3" model), we could again obtain roughly 4% performance improvement. The best performance is achieved by "7+4" which combines two CRF models and the basic Standard Matching model and the Rule Templates model. It seems that though standard matching can easily handle the correct answer, rule templates is good at dealing with semantic pattern and human naming regulation, and conditional random field can fully utilize the potential sequence information.

Figure 3 shows a case study for comparing different methods for product name recognition. The text highlighted with dark blue indicates a correct recognition by a method and text highlighted with red color indicates a wrong recognition. The standard matching can only identify those mentions occurred in the training data, while it cannot recognize many new names or new products (e.g., "lc52le810un"). The rule-based method seems a bit better by generalizing the recognition capacity using rules. However, it is still limited by the definition of rules. CRF trains a machine learning model, which is based on the quality of training data. The presented Blending model can combine the advantages of different methods, thus achieve the best performance.

Finally, one interesting point is that the performance between public and private leaderboard have huge difference, which might be resultant of the size of different training data. After investigating our final submissions, we find that we misclassify several users' names as the product mentions. We will study this problem in future work.

V. Conclusion

In this paper, we introduce our solution for the ICDM Consumer PRODucts Contest. We introduce three basic models (including standard matching, rule templates, and conditional random field) for dealing with the problem and then a blending method is proposed to combining the power of different models. The combination model achieves the best performance on Consumer PRODucts contest #1 data sets.

References

[1] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Linguisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[2] A. McCallum, "Mallet: A machine learning for language toolkit," 2002, http://mallet.cs.umass.edu.

[3] J. R. Finkel, T. Grenager, and C. D. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 2005, pp. 363–370.

[4] M. Faruqui and S. Padó, "Training and evaluating a german named entity recognizer with semantic generalization," in *Proceedings of KONVENS 2010*, 2010.

[5] M. Collins, "Ranking algorithms for named-entity extraction: boosting and the voted perceptron," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02)*, 2002, pp. 489–496.