An Overview of the CPROD1 Contest on Consumer Product Recognition within User Generated Postings and Normalization against a Large Product Catalog

Gabor Melli, Christian Romming
VigLink Inc., San Francisco, CA, USA
{gabor, christian}@viglink.com

A significant proportion of web content and its usage is due to the discussion-of and research-into consumer products. Currently however no benchmark dataset exists to evaluate the performance of text mining systems that can accurately identify and disambiguate product entities within a large product catalog. This paper presents an overview of the CPROD1 text mining contest which ran from July 2nd to Sept. 24th 2012 as part of the 21st International Data Mining Conference (ICDM-2012) that addressed this gap.

named entity recognition; named entity normalization; consumer product catalog; user generated content; CPROD1 contest

I. Introduction

A significant proportion of web usage relates to the discussions, research, and purchase of consumer products. Hundreds of thousands of blogs, forums, product review sites and e-commerce merchants currently publish information on consumer products, and a growing number of consumers use the Web to locate information on products.

This paper presents an overview of the CPROD1 text mining contest that ran from July 2nd to Sept. 24th 2012 as part of the 21st International Data Mining Conference (ICDM-2012).

The contest required that contestants develop system that can automatically recognize mentions of consumer products in previously unseen user generated web content, and to link each mention to the corresponding set of products in a large product catalog. The contest datasets includes hundreds of thousands of text-items, a product catalog with over fifteen million products, and hundreds of manually annotated product mentions to support data-driven approaches.

A high-level goal of the competition was to better understand which types of solutions can achieve winning performance on such a task. To incentivize participation we offered a prize pool of $10,000 ($6,000 for first, $3,000 for second and $1,000 for third).

The remainder of the paper is structured as follows: we begin with the rules and data files that contestants were given. Secondly, we present the annotation and data separation process. Next we describe the evaluation metric and how contestants performed during the two rounds of testing. Finally we review related benchmark tasks, and present preliminary observations about the top submitted solutions.

II. Contest Rules

Beyond Kaggle's standard terms and conditions [1] participants were requested to constrain their solutions in the following manner:

1. Participants were allowed to use additional data sources beyond the data provided by the contest, so long as: the data was publicly available, and the data was not manually transformed, such as by creating additional annotated content. If the data was based on a large Web-crawl then we required that they included the corresponding crawler code and statistics of the resulting extract.
2. Participant were required to provide the following prior to the release of the winner-selection evaluation set: 1) a trained model, 2) any additional dataset(s) used, 3) the source code and documentation required to produce predictions using their model and additional dataset(s).

III. Data Files

The CPROD1 competition involved the release of six data files[2]. Five of the files are provided immediately, while the model evaluation text-items were released near the end of the contest determine the contest winners. Files are in two formats: JSON format and .CSV format. The six files are as follows[3]:

leaderboard-text.json: This JSON file contains the text-items that participants must disambiguate to determine their leaderboard score.

products.json: This JSON file contains the product catalog that must be matched against.

training-annotated-text.json: This JSON file contains the text-items that were manually reviewed for product mentions. This file, along with training-disambiguated-product-mentions.csv, could be used to train a supervised model.

training-disambiguated-product-mentions.csv This CSV file contains disambiguated product mentions. This file, along with the training-annotated-text.json file could be used to train a supervised model. This file was in the same format as the required solutions.

training-non-annotated-text.json: This JSON file contains supplementary text-item data drawn from the same

[1] http://www.kaggle.com/terms
[2] Files can be downloaded from kaggle.com/c/cprod1/
[3] Additional file statistics can be found at kaggle.com/c/cprod1/forums/t/2199/a-first-look-at-the-data

domain as the other text-items in the contest. They are provided for participants who may opt to produce semi-supervised models.

evaluation-text.json This JSON file was provided near the end of the contest. It contained the text-item to be annotated for the final submission

Below we describe the key data entities involved (text-items, products, and disambiguated product mentions), along with the process used to generate the data.

A. *Text-items*

For the contest, a "text-item" stands for a tokenized representation of a portion or entirety of a web page or a web-forum postings page[4]. We processed each web item to create text-items as follows:

- Each text-item was automatically stripped of any HTML markup found in the source web content.
- Next, all sentences and paragraphs were automatically detected and the following special tokens inserted there: <s> and <p>. This step was known to be imprecise.
- Finally, each text-item was automatically "tokenized", where a "token" aims to separate words and other linguistic symbols used in written text, such as: punctuation marks, possessives, brackets, and quotes. This step was known to be imprecise.

Here was an example of a text-item: `"TextItem": {"0c1edc5b2ed5abb25e25b966ccdb01d2": ["Here","'s","an","example","of","a","(","pre-tokenized",")","text","item",".","<s>","<p>","Check","out","the","new","iPhone","4s","!"]}`

Notice: 1) how the word "Here's" has been divided into two tokens: "Here" and "'s"; 2) how the end-of-sentence punctuation have been placed into their own tokens; and 3) how sentences have been separated by both a "<s>" token representing an end-of-sentence and a "<p>" token representing an end of paragraph.

B. *Product-items*

For the contest a "product-item" was a semi-structured record that represents some purchasable consumer product from either the consumer electronics (CE) or automotive (AU) categories. Each record has: 1) a unique string-based identifier, and 2) an array composed of a string-based "name", a two character-based product category, and a two digit-based price. A sample of some of the products records was presented below (in tabular format):

Table 1 – Sample Product-Items

2258624	Apple LED Cinema Display 24-Inch MB382LL/A	$619.94	CE
3828027	Viewsonic's VA2448M-LED 24-Inch Widescreen LED Monitor - Black	$174.99	CE
8742810	Apple Cinema 24" Widescreen LCD Monitor - Black	$589.00	CE

[4] A forum page can contain many postings. Textitems are intended to capture a single such posting.

C. *Disambiguated Product Mentions*

For the contest a *disambiguated product mention* is a structured record composed of two fields: a product mention identifier, and a space-separated set of product-item identifiers. The product mention identifier represents some specific product mention within some specific text-item, for example: 0c1edc5b2ed5abb25e25b966ccdb01d2:0-2 represents the first through third tokens in text-item 0c1edc5b2ed5abb25e25b966ccdb01d2. The set of space separated product identifiers represents the products within the product catalog that refer to the same product as the mention.

IV. Data Preparation Process

We used the following process to annotate text-items. The annotation task involved two phases: 1) the identification of the span of tokens within text-items that identify product mention, and 2) the labeling of product items for each annotated product mention with True/False label.

During the first phase a set of text-items were randomly selected. Each text-item was reviewed by at least two different annotators. In cases where there was disagreement about mentions a third annotator broke ties.

During the second phase the human annotators were asked to classify which products were legitimate references for each of the product mentions. This phase was significantly more time consuming so only a small portion of product candidates were reviewed by two or more annotators.

Finally, we randomly separated the annotated text-items into training set, leaderboard set, and model evaluation set using the following proportions: 50%, 25% and 25%.

V. Evalution Metric

Each submission of annotations was scored based on the average F1 score (between 0.0 and 1.0) for the union of predicted and true disambiguated product mentions. The table below illustrates the performance calculation for a single participant who scored 0.414. Table 2 illustrates all possible outcomes for a prediction set. In this scenario the participant submitted six disambiguated product mentions (pm_1 … pm_6) while the truth set contained six manually annotated product mentions that were hidden from the participant (tm_1 … tm_6). Notice that one of the predicted mentions, pm_6, was not in the truth set (their start and end tokens do not align) and one of the mentions in the truth set, tm_3, was not in the predicted set. Each of these outcomes is assigned an F1 score of 0. The remaining five predictions can be scored based on the F1 calculation based on the predicted products.

VI. Model training Phase

From July 2nd through Sept 15th teams were able to submit predictions against the leaderboard-text.json file in order to evaluate their performance and determine their ranking. The table below shows the final scores and rankings along with the number of submissions by each team.

Table 2 – illustration of the evaluation on a set of predictions and actual mentions.

Predicted Mention	True Mention	Predicted Product	True Product	Correctness (TP,FP, FN)	P	R	F1
pm_1	tm_1	484946	484946	TP	1.0	1.0	1.0
pm_2	tm_2	0	0	TP	1.0	1.0	1.0
not predicted	tm_3	not predicted	103492	FP	0.0	0.0	0.0
pm_3	tm_4	Not predicted	0	FN	0.0	0.0	0.0
		223801	not in actual	FP			
pm_4	tm_5	167712	167712	TP	0.5	0.5	0.5
		Not predicted	385994	FN			
		194730	not in actual	FP			
pm_5	tm_6	250747	250747	TP	0.5	0.33	0.4
		Not predicted	237004	FN			
		Not predicted	482721	FN			
		722416	not in actual	FP			
pm_6	Not a mention	416094	not in actual	FP	0.0	0.0	0.0
						avg(F_1)=	0.41

Table 3 summarizes the results of the leaderboard phase. The eight teams that also submitted predictions against the evaluation-text.json file are in bold casing. The first baseline was based on a dictionary lookup of terms found in the training data. Any mention of the term in the leaderboard file was deemed to be a product mention associated with the same products. The second baseline was based on CRF trained model. This baseline performed worse than the first baseline because it always predicted the null product (0) for each mention.

VII. Winner Selection

Eight teams continued to the final and deciding phase of the competition by submitting their predictive systems by Sept. 15th and also predictions against the evaluation-text.json file that was released on Sept. 16th. Table 4 summarizes the F1 score of all eight participants – with the three winning teams in bold case.

Table 3 – Final Leaderboard Results

rank	team	score	submissions
1	**8000**	0.31495	48
2	**ISSSID**	0.30656	19
3	**mt.banahaw**	0.26644	10
4	**SINGA**	0.24816	55
5	**Balazs Godeny**	0.22137	14
6	student2012	0.22009	10
7	**Olexandr Topchylo**	0.21425	21
8	tuzzeg	0.18232	9
9	**Labeler**	0.18229	2
10	Vivek Sharma	0.17262	6
11	seemla	0.16667	2
11	Nikit Saraf	0.16667	7
13	**dvg**	0.15385	10
14	Pan Kidd	0.15287	7
15	NDB	0.14047	6
	baseline1	0.12500	
16	LCCK	0.12500	2
16	brotherC	0.12500	1
19	Roberto-UCIIIM	0.10922	3
20	3john	0.09954	10
21	dmcat	0.08454	8
22	Bart	0.07809	4
	baseline2	0.07471	
23	Navin K	0.07471	2
25	M. Hu	0.07344	4
26	SmallAnt	0.03595	4
27	Jacques Kvam	0.00643	2
28	Trung Huynh	0.00022	1
29	FL	-	2
30	AdjustedRSquared	-	3

The baseline 1 entry was based on the same dictionary-lookup mechanism used in the leaderboard baseline. The lower score of the baseline on the evaluation set relative to the leaderboard (0.093 vs. 0.125) suggests that the final evaluation was more difficult.

Table 4 – Final Winning Results

Rank	Team	F1 Score
1	**ISSSID**	0.22041
2	**Olexandr Topchylo**	0.19883
3	**8000**	0.18780
4	Balazs Godeny	0.18778
5	Labeler	0.16444
6	SINGA	0.16037
7	mt.banahaw	0.13003
8	dvg	0.10593
	baseline 1	0.09302

VIII. Related Contests and Datasets

Several datasets exists that resemble CPROD1's. Two of them are itemized below: TAC 2012 and BioCreative III. Our task differs from them, in the following aspects:

a) the token span of consumer product mentions are more ambiguous than people, places or protein mentions.

b) often, product mentions differ substantially from how they appear in a catalog

d) the consumer product domain is a very important and special domain by itself. However, we've seen very few standard benchmark datasets available for further research.

D. *TAC 2010 Entity-Linking Task*

The TAC 2010 entity-linking task [1] [5] requires that participants, given 1) a name/term (of a Person, Organization, or Geopolitical Entity), 2) a document (in English, Chinese, and Spanish) containing that name, and 3) a knowledge-base derived from English Wikipedia, must determine the KB node for the named entity, adding a new node for the entity if it was not already in the KB. This task was similar to our entity-linking task. TAC 2012 however continues with the use of commonly targeted entity types of person, organization and geopolitical entity. No prize money was awarded.

E. *BioCreative III Gene Normalization Task*

The BioCreative III Gene Normalization Task [2] [6] requires that participants link gene or proteins mentioned in the literature to standard database identifiers (EntrezGene Ids). The task was similar to our entity-linking task. BioCreative III however was aimed at bioinformatics. No prize money was awarded.

IX. Observations

Below are some of the preliminary observations about the final submissions

- Several of them were able to achieve competing performance based on gazetteer-lookup solutions.
- Several to them trained sequential tagging models to the recognition task.
- None of them appeared to use a semi-supervised approach on the non-annotated data.
- None of them appeared to train a model to perform normalization.

Acknowledgment

We would like to thank the following people for their assistance in delivering the contest. First, we would like to thank Wei Fan, Geoff Webb, Bart Goethals and Jilles Vreeken from the ICDM 2012 organization team for constructive feedback in the contest structure and assistance throughout. Next we thank the VigLink engineering team. Edward Chu and Rich Fletcher created our critical annotation environment, and Nam Pham and Idris Raja who assisted us in the paper reviews. From Kaggle we thank Joyce Noah-Vanhoucke for assistance throughout the contest and particularly throughout the challenge of implementing the evaluation metric. We would also like to thank Yabo (Arber) Xu for helping us to promote the contest in Asia. We, of course, are very thankful to the annotators who persevered for all those hours to create the CPROD1 dataset. Finally, we are thankful to all the participants who attempted the challenge!

References

[1] H. Ji, R. Grishman, and H.T. Dang, "Overview of the TAC 2010 Knowledge Base Population Track," In TAC (Text Analysis Conference 2010 Workshop.

[2] Z. Lu, H.-Y. Kao, et al. "The Gene Normalization Task in BioCreative III." In: BMC Bioinformatics, 12(8), 2011.

[5] www.nist.gov/tac/2012/KBP/TAC_2012_KBP_CFP.pdf

[6] www.biocreative.org/tasks/biocreative-iii/gn/