SINAI at WEPS-3: Online Reputation Management

M.A. García-Cumbreras, M. García-Vega
F. Martínez-Santiago and J.M. Peréa-Ortega
University of Jaén. Departamento de Informática
Grupo Sistemas Inteligentes de Acceso a la Información
Campus Las Lagunillas, Ed. A3, E-23071, Jaén, Spain
{magc, mgarcia, dofer, jmperea}@ujaen.es

Abstract

The online reputation management systems help to the consumers to make buying decisions looking for opinions in the web about many products offered by companies, also interested in the same opinions. This paper presents the system developed by the SINAI research group at the WEPS-3 task, called Online Reputation Management.

Given a Twitter entry and a company name, the goal is to decide if the entry talks about this company. Our system is based on the use of linguistic information in Twitters entries for extracting information and creating an XML data collection about companies referred in the Twitter entries. This XML collection is filled extracting information from Internet web pages, like Wikipedia, and the use of the DBpedia ontology.

Using this collection and some logical rules we have developed a promising system, with a very good precision and easily extensible with the addition of new rules and other Internet resources.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms

Algorithms, Experimentation, Languages, Performance

Keywords

On-line Reputacion Management, Opinion Mining, Linguistic Rules

1 Introduction

Consumers use the web to make buying decisions. A majority buys goes online to research, read reviews and get opinions from other consumers. With the growth of consumer-generated media (CGM) such as blogs, forums, and message boards, information can be quickly generated and indexed by search engines.

Even after many years of clean business practices, a single negative event can stain your brand image in the public eye for a long time. A negative product review or a negative comment in social

blogs like Twitter or Facebook can be detrimental to your brand, especially when competitors are standing close by to snatch up customers. One way to combat that threat is through a reputation management strategy with the so called On-line Reputation Management (ORM) systems.

A first step in a ORM system is the detection of the company given a set of opinions. The ambiguity of names is an important bottleneck for the experts who want to analize these opinions.

This paper presents the system developed by the SINAI research group at the WEPS-3 task, the ORM task.

The main goal of this task is to discover if a Twitter entry belongs to a given company. As the organization said, Twitter has been chosen as target data because it is a critical source for real time reputation management and also because ambiguity resolution is challenging: tweets are minimal and little context is available for resolving name ambiguity.

Existing ORM systems[2] use, on one hand, machine learning techniques and features extraction to train a system and the test entries are tagged with the model trained. On the other hand, manual rules can be applied to decide if an entry talks about a company or not.

Our system works with linguistic rules[1] with the aim to work in real time. Based on these rules, training data are no neccesary nor features extraction.

The following section describes the system developed. In Section 3, we explain the experiments and obtained results. Finally, conclusions are presented in Section 4.

2 Description of the system

A Twitter entry may contain multiple entities and proper nouns that refers to a company. Given an entry and a company name, our objective is to determine the company that each entity or proper noun refers to decide if the entry talks about this company. We use some linguistic rules to perform this task.

The first module of the system analize the Twitter entry, extracting each pair entity-type found. The Name Entity Recognizer (NER) used was the module included in GATE[1]. GATE (General Architecture for Text Engineering) is a stable, robust, and scalable infrastructure which allows users to build and customise language processing components.

The second module create for each organization a complete XML file with the following data:

- Organization name.
- Webhome. The organization of WePS-3 delivered a file with all the organizations, where each organization appears with its webhome and several data. The webhome was extracted and preprocessed (html tags and stopwords removal).
- Wikipedia page. For each company the data of its wikipedia[2] page, if exists, has been extracted and added to its XML file.
- DBpedia abstract. The DBpedia Ontology[3] is a shallow, cross-domain ontology, which covers over 259 classes, described by 1,200 different properties, forming a subsumption hierarchy. For each company the data of the abstract property, if exists, has been added to the XML file. The abstract property in DBpedia describes the company and its main features.
- DBpedia products. For each company the data of the products property, if exists, has been added to the XML file. The products property includes the products that this company manufactures.

The following text shows an example of the XML file created for each company.

[1] available at http://gate.ac.uk/
[2] available at http://www.wikipedia.org/
[3] available at http://dbpedia.org/

```
<doc>
  <field name="id">Starbucks</field>
  <field name="webhome" url="http://www.starbucks.com/">Starbucks Coffee Company
 skip to Main Navigation...</field>
  <field name="DBpediaAbstract">Starbucks Corporation is an international coffee
 and coffeehouse...</field>
  <field name="DBpediaProducts">Merchandise Baked goods Smoothies Frappuccino
 beverages Bottled beverages Made to order beverages Boxed tea Whole bean
coffee Alcoholic Beverages</field>
</doc>
```

The third module applied the manual rules generated. These rules are the following:

- (sinai_1) The name of the company appears in the Twitter entry.
- (sinai_2) The name of the company appears in the entities detected in the Twitter entry, and the entity recognized is the type "organization".
- (sinai_3) At least one the entities detected in the Twitter entry appears in the organization webhome.
- (sinai_4) At least one the entities detected in the Twitter entry appears in the wikipedia page of the organization.
- (sinai_5) At least one the entities detected in the Twitter entry appears in the DBpedia abstract or DBpedia products of the organization.

The result of this third module for each Twitter entry and its possible organization is a TRUE/FALSE value, that confirm if the entry belongs to this organization.

3 Experiment Description and Results

We carried out experiments using the framework given by the WePS-3 organizers. The data set contains 47 organizations, with roughly 500 Twitter entries for each organization, 22490 entries in total.

Applying the linguistic rules described, most of the tags are true. The distribution of tags obtained is presented in the Table 1. The sum of the tagged entries is 22013 because one of the organizations cause an error in the system and its entries were not tagged.

Experiment	True	False
sinai_1	6701	15312
sinai_2	0	22013
sinai_3	15758	6255
sinai_4	5528	16485
sinai_5	11548	10465

Table 1: Distribution of true/false tags

Table 2 presents the results obtained with these five runs. Second column shows the accuracy value (Acc), and the following columns present the precision (P) and recall (R) obtained with every experiment, with the positive (pos) and negative (neg) values. Last column shows the ranking value obtained with the other systems presented in this WePS-3 task.

After the analysis of the results of these fiv experiment we conclude:

Experiment	Acc	P(pos)	R(pos)	P(neg)	R(neg)	Rank
sinai_1	**0,63**	0,84	0,37	0,68	0,71	6/18
sinai_2	0,56	1	0	0,58	0,98	9/18
sinai_3	0,46	0,6	0,7	0,86	0,28	13/18
sinai_4	0,61	**0,9**	0,26	0,73	0,72	7/18
sinai_5	0,51	0,72	0,51	0,75	0,47	11/18

Table 2: SINAI results for the five experiments with linguistic rules

1. The best result has been obtained with the first experiment (sinai_1), with a 0,63 value of accurary. Almost the 80% of the tags were tagged as FALSE, and the precision value for positive tags achieves a 0,84, so this simple rule works well for the positive tags, and not bad for negative tags.

2. The second experiment (sinai_2) tagged all the entries as FALSE, because the NER module did not work well to classify the entities, because of the length and format of the Twitter entries. The result obtained is not relevant.

3. Webhome information contains general text with no relevant information in the most of cases, so a lot of the entries were tagged as TRUE, and the results were poor. Only we can emphasize that the recall for positive tags achieves a value of 0,7.

4. Wikipedia contains relevant and summarized information about a company. This information was not available for all the organizations, and in some cases if the organization name is also a common word, like apple, obtained a wrong Wikipedia webpage. The result was good in terms of accuracy, 0,61, and also in terms of precision for positive and negative. We emphasize the recall value obtained for negative tags, 0,72.

5. The results obtained using DBpedia information (abstract and products) are not relevant. Only for a low percentage of the organization we oatained this information (in some cases our system did not find this information or the name of the organization in DBpedia was different than our query). In other cases, if the organization name is also a common word, like apple, the information obtained was wrong.

4 Conclusions

In this paper we have presented the experiments carried out for the CLEF 2010 WePS task (Online Reputation Management). We conducted diagnostic experiments applying different logical rules over the data offered by CLEF. This rules used an XML information collection built from the Wikipedia y DBpedia and the behavior of the system has been very promising. We have obtained a 0.63 value of accuracy of the best rule, obtaining a precision of 0.84 for the positive tags and 0.72 for the negatives.

Nevertheless, the nature of task and the training data give us a guide to improve the results. The absence of information for some companies in Wikipedia and the low relevance of DBpedia encourage us in the use of more specific information sources in Internet. Like the information offered by the sources used is so sparse, we think that a combined system will improve the precision, e.g. weighting the rules using some kind of learning algorithm based on the information given of each company.

Acknowledgements

This work has been partially supported by a grant from the Spanish Government, project TEXT-COOL 2.0 (TIN2009-13391-C04-02), a grant from the Andalusian Government, project GeOasis

(P08-TIC-41999), and a grant from the University of Jaen, project RFC/PP2008/UJA-08-16-14 and project UJA2009/12/14.

References

[1] Xiaowen Ding and Bing Liu. The utility of linguistic rules in opinion mining. In Wessel Kraaij, Arjen P. de Vries, Charles L. A. Clarke, Norbert Fuhr, and Noriko Kando, editors, *SIGIR*, pages 811–812. ACM, 2007.

[2] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.