

WePS-3 Evaluation Campaign: Overview of the Online Reputation Management Task

Enrique Amigó¹, Javier Artiles¹, Julio Gonzalo¹, Damiano Spina¹, Bing Liu²,
and Adolfo Corujo³

¹ NLP Group of UNED University,
Madrid, Spain.

<http://nlp.uned.es>

² Department of Computer Science,
University of Illinois at Chicago, USA.

<http://www.cs.uic.edu>

³ Lorente & Cuenca, Communication Consultants
Madrid, Spain.

<http://www.llorenteycuenca.com>

Abstract. This paper summarizes the definition, resources, evaluation methodology and metrics, participation and comparative results for the second task of the WEPS-3 evaluation campaign. The so-called Online-Reputation Management task consists of filtering Twitter posts containing a given company name depending of whether the post is actually related with the company or not. Five research groups submitted results for the task.

1 Introduction

People share opinions about products, people and organizations by means of web sites such as blogs, social networks and product comparison sites [8, 6]. Online reputation management (ORM) consists of monitoring media, detecting relevant contents, analyzing what people say about an entity and, if necessary, interact with costumers. Negative comments in online media can seriously affect the reputation of a company, and therefore online reputation management is an increasingly important area of corporate communication.

Perhaps the most important bottleneck for reputation management experts is the ambiguity of entity names. For instance, a popular brand requires monitoring hundreds of relevant blog posts and tweets per day; when the entity name is ambiguous, filtering out spurious name matches is essential to keep the task manageable.

WePS-3 ORM task consists of automatically filter out tweets that do not refer to a certain company. In particular, we focus on the Twitter social network because (a) it is a critical source for real time reputation management and (b) also because ambiguity resolution is particularly challenging: tweets are minimal and little context is available for resolving name ambiguity.

This task is a natural extension of WePS evaluation campaigns, which have been previously focused on person name ambiguity in Web Search results; with the ORM task, WePS-3 extends its scope to cover other relevant type of named entity. Our task is related to the TREC Blog Track [7], which focused on blog posts. However, in that case, systems dealt with information needs expressed by queries, rather than focusing on a name disambiguation problem.

2 Task definition

2.1 Twitter

Twitter is a relatively new social networking site [4] referred to as a microblogging service. Its particularity is that posts do not exceed 140 characters and there are no privacy conditions. Therefore, Twitter reflects opinions in real time and it is very sensitive to burstiness phenomena.

Tweets are particularly challenging for disambiguation tasks given that the ambiguity must be sorted out using a very small textual context.

2.2 Ambiguity

The idea of ambiguity is actually quite fuzzy. For instance, suppose that we are interested in a certain car brand. If the brand name is common, of course, occurrences that refer the common word sense are not related to the brand. But let us suppose that the brand sponsors a football team. We could think that the referred organization is actually the football team, but not the brand. But experts could be interested on monitoring these occurrence given that they have spend money to be mentioned in this way. In addition, experts might be interested on mentions to the brand generically, but not on specific products (which might be handled separately). In short, the ambiguity is closely related with the concept of relevance, which is inherently fuzzy.

For evaluation purposes, one option consists of defining the relevance criteria for each entity just like in other competitive tasks as TREC [7]. However, interpreting the relevance criteria can be difficult even for humans. Probably, systems will not able to tackle this issue. Indeed, interpreting the relevance criteria can be difficult even for humans.

In this competition we opt for a lax interpretation of relevance, considering ambiguity at a lexical level: the sense of the name must be derived from the company, even if the sentence does not explicitly talk about the company. Table 1 illustrates this idea for the Apple company.

2.3 Input and output data

The first decision when defining system input and output is whether systems should be able to use a training set for each of the companies included in the test set. There are two possible scenarios: an ORM company that provides individualized services to a limited number of clients, or an online system that

...you can install 3rd-party apps that haven't been approved by Apple..	TRUE
...RUMOR: Apple Tablet to Have Webcam, 3G...	TRUE
...featuring me on vocals: http://itunes.apple.com/us/album/...	TRUE
...Snack Attack: Warm Apple Toast...	FALSE
...okay maybe i shouldn't have made that apple crumble...	FALSE

Table 1. Examples of tweet disambiguation for the company Apple

accepts any company name as input. In the first scenario, the system will probably be trained for each of the clients. In the second scenario, this is not viable, as the system must immediately react to any imaginable company name. We have decided to focus on the second scenario, which is obviously the most challenging. Therefore, the set of organization names in the training and test corpora are different.

For each organization in the dataset, systems are provided with the company name and its homepage URL. This web page contains textual information that allows systems to model the vocabulary associated to the company. The input information per tweet consists of a tuple containing: the tweet identifier, the entity name, the query used to retrieve the tweet, the author identifier and the tweet content.

3 Data set

3.1 Trial Corpus

The trial corpus consists of 100 tweets per organization name. 24 companies were selected; 18 from English speaking countries and 6 from Spanish speaking countries. Most of these entities were extracted from a Twitter Brand Index that appears in the blog “Fluent Simplicity”⁴. Table 2 enumerates these entities and the category associated in the brand index.

The first observation was that identifying companies for our purposes was not a trivial task. The first reason is that many companies are not usually mentioned in Twitter. Tweets tend to focus on certain issues. For instance, some frequent issues are entertainment technologies, movies, travel, politics, etc. Therefore, most companies do not have enough presence in Twitter to be included in our test bed. In addition, many company names are either too ambiguous or not ambiguous at all. For instance, “British Airways” is not ambiguous. However, in order to ensure a high recall, we should use the query term “British” (e.g. “I fly with British”). But in this case, 100 tweets would not be enough to obtain true samples. Notice that this does not imply that our systems would not be useful to monitor British Airways. The key issue is that we need reasonably ambiguous company names in order to make the annotation task feasible. In short, the company selection is very costly, given that it requires to retrieve and check tweets manually to analyze their ambiguity.

⁴ <http://blog.fluentsimplicity.com/twitter-brand-index/>

Entity name	Query	Language	Category
Best Buy	best buy	English	Online-shopping
Leap frog	leapfrog	English	toys
Overstock	overstock	English	Online-shopping
Palm	palm	English	Mobile products
Lennar	lennar	English	home builder
Opera	opera	English	Sofstware
Research in motion	rim	English	Mobile products
TAM airlines	tam	English	Airline
Warner Bros	warner	English	Films
Southwest Arilines	southwest	English	Airline
Dunkin Donuts	dunkin	English	Food
Delta Airlines	delta	English	Airline
CME group	cme	English	Financial group
Borders bookstore	borders	English	bookstore
Ford Motor	ford	English	Motor
Sprint	sprint	English	Mobile products
GAP	gap	English	Clothing store
El hormiguero	hormiguero	Spanish	TV program
Renfe Cercanas	cercanias	Spanish	commuter train service
El Pas	pais	Spanish	Newspaper
El Pozo	pozo	Spanish	Food
Real madrid	madrid	Spanish	Soccer team
Cuatro	cuatro	Spanish	TV chanel

Table 2. Selected tweets for trial corpus

For each company, the first 100 tweets retrieved by the corresponding query have been annotated directly by the task organizers. During the annotation, we observed that the best approach consisted of detecting key terms associated to the company. In some cases these key terms were related with a certain event that happens just before the retrieval process (such as, for instance, a new product launched by Palm).

3.2 Training and test corpus

The initial purpose was to define a methodology for the company selection. Fluent Simplicity was not enough to pick them. The next attempt consisted of filtering automatically the companies included in DBpedia ⁵ which is a knowledge base that extracts structured information from Wikipedia. The automatic filter consisted of detecting company names that match common names. This should ensure the ambiguity of names. However, the presence in Twitter was less frequent than companies from the Twitter Brand Index. In addition, again, some company names were either too much ambiguous or not ambiguous at all. Finally, the list was expanded with a few entities that are not exactly a company, such as sport teams or music bands, which are very common in Twitter.

⁵ <http://dbpedia.org/About>

Although the original plan was to annotate around 500 entities, the training and test corpus finally contains 100 company names. We have discarded Spanish companies given that, for now, Twitter is still far less popular than in English speaking countries. Table 5 shows the entities selected for the training and test corpora.

4 Assessments

4.1 Mechanical Turk

The training and test corpora have been annotated by means of Mechanical Turk services. The advantages of using this service for annotation have been reported in previous work [5, 3] Figure 1 shows an example of our formularies for Mechanical Turk. Each hit contains five tweets from the same company name to be annotated. It also includes a brief description for the company and the annotator can access the company web page. In order to ensure that tweets have been annotated, there is no default value for the annotation. The annotation options for each tweet were “related”, “non related” or “undecidable”. Each hit has been redundantly annotated by five Mechanical Turk workers. The form includes the following instructions to annotators:

The next table contains tweets that apparently mention a company. The task consists of determining whether each tweet mentions the company (button “related”), does not mention the company (button “non related”) or there is not enough information to decide it (button “undecidable”). This page provides the company name and its URLs. For each tweet the table includes the tweet author and content. Notice that most tweets contain links that can help you make this decision. Find below some examples for the Apple company.

902 annotators participated in the annotation of 43730 tweets. Given that not all company names had the same presence in Twitter and some tweets have been discarded, the number of annotated tweets per entity is variable; between 334 and 465 tweets.

4.2 Agreement analysis

Figure 2 shows the relationship between the average agreement for single turkers versus the number of annotated tweets. The averaged agreement of single turkers is computed as the number of annotators that have taken the same decision (related, non related or undecidable). As the figure shows, most annotators have an average agreement with other annotators between 3 and 4.5. That means that in most cases at least 3 annotators have taken the same decision.

4.3 Ground truth

We have followed the following criteria to decide the final annotation (related or non related) for each tweet:

Company name: Apple

Apple Inc. is an American multinational corporation that designs and manufactures consumer electronics, computer software, and personal computers.

Home Page: <http://www.apple.com>

author	tweet text	assessment
johny12	you can install 3rd-party apps that haven't been approved by Apple	related non related undecidable
charlotte.jones	RUMOR: Apple Tablet to Have Webcam, 3G	related non related undecidable
mike46	featuring me on vocals: http://itunes.apple.com/us/album/	related non related undecidable
ryan13	Snack Attack: Warm Apple Toast	related non related undecidable
hallmorton	okay maybe i shouldn't have made that apple crumble	related non related undecidable
jjones12	Apple is a good choice.	related non related undecidable

Fig. 1. Example of form for Mechanical Turk annotation

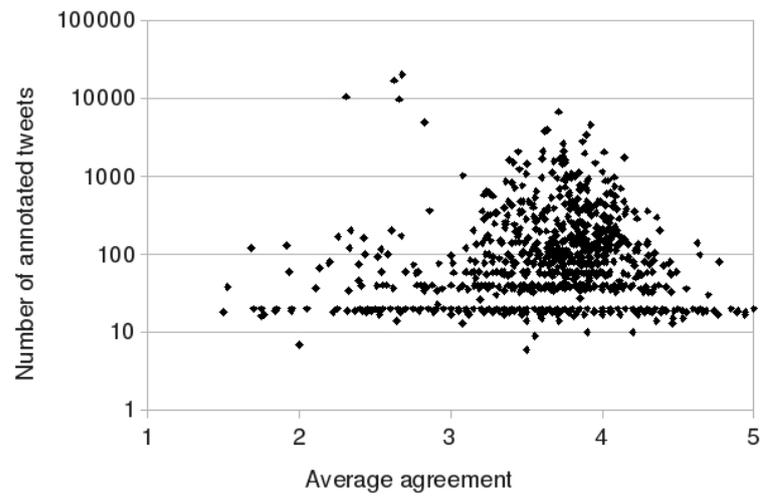


Fig. 2. Average agreement for single turkers versus number of annotated tweets.

- If four or five annotators take the same decision, then this corresponds with the ground truth. This set represents 58% of tweets.
- If three or more annotators agree and there is no more than one disagreeing annotator, then we also consider that it is the ground truth. We consider that two annotators disagree when one says “related” and the other says “non related”. This sample set represents the 21% of cases.
- The most controversial case is when three annotators are contradicted by two annotators. These are 14% of cases. We analyzed manually 100 samples and we found that the three votes corresponded with the ground truth in around 80% of cases. At the risk of introducing a bit of noise in the corpus, we have considered the majority of votes as the ground truth. In any case, system evaluation results did not change substantially when considering these cases.
- In a 0.1% of cases, there were less than 2 related and non related votes, in favor of undecidable votes. We have directly discarded these cases.
- In 7% of cases there were two related votes and two, related votes and one undecidable. These cases have been meta-evaluated manually by the task organizers.

4.4 Entity ambiguity

Figure 3 shows the distribution of ambiguity across company names. That is, the ratio of related tweets for each entity. The company names have been sorted according to their ratio. As the figure shows, although we have tried to select names with medium ambiguity, there is a great variability of ambiguity in the corpus and there is an important amount of companies with low occurrence in tweets. This has important implications in the evaluation metric definition. It is desirable to check to what extent systems are able to detect the ratio of related tweets for each single company name.

5 Evaluation metrics

Basically, this task can be considered as a classification task. The most natural way of evaluating is the accuracy measure. That is, in how many cases the system output matches the annotation. However, this metric does not consider the distribution of related and non related tweets within the correct outputs. That is, for a high ambiguous company name, even only a few related tweets appeared in the corpus, the decisions taken in these cases are crucial. This issue has relevance in this corpus given that most of company names have a very high or low ambiguity. We consider this aspect by computing also the precision and recall over both the related and non related classes. In addition, the F measure of precision and recall is computed for each company name and class.

Another important aspect is how to consider the cases in which the system does not return any results. In the case of accuracy, these are fails. In term of precision and recall measures, these cases affect by decreasing the recall for the corresponding class (related or non related).

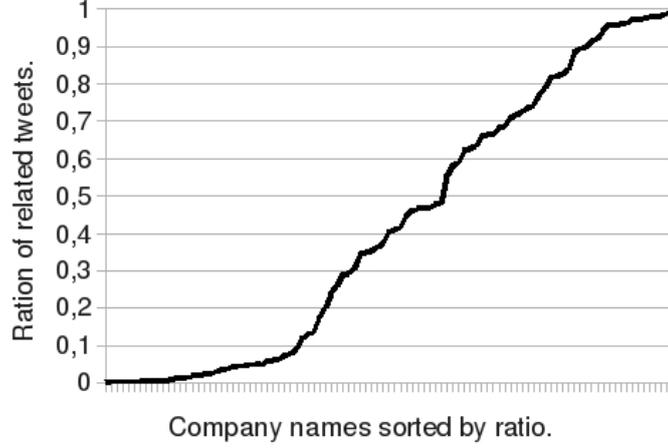


Fig. 3. Ambiguity distribution across company names

Finally, we are interested in knowing to what extent the systems are able to predict the ratio of related tweets given a query. It is important because this ratio is enough to estimate the entity popularity in Twitter. In theory, estimating this ratio does not strictly require to know what tweets are related and what not. We define the Related Ratio Deviation as the absolute difference between the real ratio and the ratio given by the system.

Considering the six categories for the sample set T : true positive (TP), False positive (FP), true negative (TN), false negative (FN), empty outputs for positive inputs (EP) and empty output for negative inputs (EN), our measures are defined as:

$$\text{Accuracy} = \frac{TN + TP}{T}$$

$$\text{Precision over the related class} = \frac{TP}{TP + FP}$$

$$\text{Recall over the related class} = \frac{TP}{TP + FN + EP}$$

$$\text{Precision over the non related class} = \frac{TN}{TN + FN}$$

$$\text{Recall over the non related class} = \frac{TN}{TN + FP + EN}$$

$$\text{Related Ratio Deviation} = \text{abs} \left(\frac{(TP + FP) - (TP + FN + ET)}{T} \right)$$

The accuracy metric assigns a relative weight to the related and non related classes depending on the distribution of both classes in each company name. That is, the more the tweets are related to the company, the more this class is

considered in the evaluation process. However, this weighting criterion is arbitrary. In addition, the combination of precision and recall measures by means of the F measure over each class assumes that both precision and recall have the same weight. The final ranking could change if we employed a different metric weighting criterion.

For this reason, for each system pair we check to what extent the improvement is robust across potential metric weighting schemes by applying the UIR measure [1]. This measure was also employed in WEPS2 campaign [2]. Being $T_{\forall m.a>b}$ the number of company names such us System a improves System b for all the four metrics, and being T the total number of company name (test cases), UIR is defined as follows:

$$UIR(a, b) = \frac{T_{\forall m.a>b} - T_{\forall m.b>a}}{T}$$

The more System b improves System a for all metrics (or there are contradictory results between metrics), the more $UIR(a, b)$ increases (decreases). We have combined the four precision and recall metrics with UIR.

6 Participation and evaluation results

16 runs have participated in the task. Table 3 shows the evaluation results sorted by accuracy. Two baseline systems have been added to the ranking, consisting of tagging all tweets as related ($Baseline_R$) or non related ($Baseline_{NR}$). The first observation is that the ranking discriminates participating groups.

The top system is LSIR-EPFL. The main particularity of this system is the use of additional resources for classification which include Wordnet, meta-data from the web page, Google results, and user feedback (just some words). Their experiments showed that even excluding the user feedback, they obtained high accuracy. According to their experiment description, using the same approach but considering just the company web page, the evaluation results would descend to the middle of the ranking.

The system SINAI (located in the middle of the ranking) also employs additional resources, but they basically consist of named entities extracted from the tweets while LSIR-EPFL employs all the tweet content. A deeper analysis showed that there is a great variability of evaluation results for this system across company names. For some company names, the system improves the top ranked system, while for other names, it achieves very low results. This variability is not related with the ratio of related tweets for the company name. Therefore it is not due to classification thresholds. In short, the SINAI evaluation results suggest that considering the named entities appearing in tweets is appropriate for certain company names.

The second best system is ITC-UT, which uses an initial classification step to predicting the ambiguity of the company name, according to some evidences. The classification step consisted of a set of rules based on Part of Speech tagging and Named Entity recognition. Given that the system variants do not differ from

Run	Non processed tweets	precision (related)	Recall (related)	F measure (related)	precision (non related)	recall (non related)	F measure (non related)	Accuracy	Related Ratio Deviation
LSIR.EPFL 1	0	0.71	0.74	0.63	0.84	0.52	0.56	0.83	0.15
ITC-UT 1	0	0.75	0.54	0.49	0.74	0.6	0.57	0.75	0.18
ITC-UT 2	0	0.74	0.62	0.51	0.74	0.49	0.47	0.73	0.23
ITC-UT 3	0	0.7	0.47	0.41	0.71	0.65	0.56	0.67	0.26
ITC-UT 4	0	0.69	0.55	0.43	0.7	0.55	0.46	0.64	0.32
SINAI 1	449	0.84	0.37	0.29	0.68	0.71	0.53	0.63	0.36
SINAI 4	449	0.9	0.26	0.17	0.73	0.72	0.53	0.61	0.38
BASELINE _{NR}	0	1	0	0	0.57	1	0.66	0.57	0.43
SINAI 2	449	1	0	0	0.58	0.98	0.65	0.56	0.43
UVA 1	409	0.47	0.41	0.36	0.6	0.64	0.55	0.56	0.27
SINAI 5	449	0.72	0.51	0.28	0.75	0.47	0.33	0.51	0.48
KALMAR R. 4	874	0.48	0.75	0.47	0.65	0.25	0.28	0.46	0.43
SINAI 3	449	0.6	0.7	0.36	0.86	0.28	0.19	0.46	0.54
KALMAR R. 2	874	0.47	0.7	0.43	0.61	0.27	0.28	0.44	0.43
KALMAR R. 5	874	0.48	0.77	0.47	0.65	0.21	0.23	0.44	0.45
BASELINE _R	0	0.43	1	0.53	1	0	0	0.43	0.56
ALMAR R. 1	2207	0.51	0.7	0.42	0.59	0.19	0.21	0.4	0.39
KALMAR R. 3	2202	0.49	0.66	0.39	0.66	0.25	0.27	0.4	0.47

Table 3. Final ranking

each other substantially, it is difficult to know what aspect lead the system to get ahead other systems. However, this result shows that it is possible to obtain an acceptable accuracy just considering linguistic aspects of the company mention.

The system UVA makes a relevant contribution to the task results. This system does not employ any resource related with the company, such as the web page or Google results. Although the accuracy results are not very high, the Related Ratio Deviation is as low as the systems located at the top of the ranking. This result suggests that a general classifier can be employed to predict the presence of any company in Twitter.

Finally, the Kalmar system employs a bootstrapping method starting from the vocabulary of the web page. The global accuracy results are not very high, but a deeper analysis shows that this approach improves the best system in terms of F measure over the related class when just a few tweets are relevant in the collection. In general, systems tend to achieve low F measure over the related class when the related class is not frequent. This does not happen in the case of Kalmar system. In other words, Kalmar results suggests that bootstrapping is appropriate for company names with high ambiguity.

Table 4 shows the UIR results. The third column represents the set of systems that are improved by the corresponding system with no dependence on metric evaluation weightings. As the table shows, the top system, in addition to achieve higher Accuracy, improves robustly most of the other systems. Of course, although a baseline system (all tweets are non related) appears in the middle of the ranking, it does not improve robustly any other system: it is just an effect of the metric combination used to rank systems.

Run	Accuracy	Improved systems
LSIR.EPFL 1	0.83	KALMAR R. 1 KALMAR R. 5 ITC-UT 2 KALMAR R. 2 KALMAR R. 3 ITC-UT 4 KALMAR R. 4 UVA 1 BASELINE _R
ITC-UT 1	0.75	SINAI 4 UVA 1
ITC-UT 2	0.73	SINAI 4, UVA 1
ITC-UT 3	0.67	KALMAR R. 2, KALMAR R. 3, UVA 1,
ITC-UT 4	0.64	SINAI 4, UVA 1
SINAI 1	0.63	SINAI 4, SINAI 2, UVA 1, BASELINE _{NR}
SINAI 4	0.61	
BASELINE _{NR}	0.57	
SINAI 2	0.56	
UVA 1	0.56	KALMAR R. 1, KALMAR R. 2, KALMAR R. 3
SINAI 5	0.51	
KALMAR R. 4	0.46	KALMAR R. 1, KALMAR R. 5
SINAI 3	0.46	
KALMAR R. 2	0.44	
KALMAR R. 5	0.44	
BASELINE _R	0.43	
KALMAR R. 1	0.4	BASELINE _{NR}
KALMAR R. 3	0.4	KALMAR R. 1

Table 4. UIR results. UIR threshold = 0.1

7 Conclusion

This competition is the first attempt to define a shared task to solve the problem of company name disambiguation in social networks (Twitter in our case). Our conclusion is that it is a task feasible to evaluate, given that we have obtained an acceptable agreement between Mechanical Turk annotators. A corpus with around 20,000 annotated tweets is now available for future benchmarking.

The evaluation results have shed some light on how to solve the task: (i) Considering additional sources like Google results or wordnet seems to be useful; (ii) linguistic aspects of the company mention are also very indicative (iv) It is possible to define a general approach to estimate approximately the presence of a company name in Twitter (v) Finally, bootstrapping methods seems to be useful, specially for highly ambiguous company names.

8 Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Innovation within the project QEAVis-Catiex (TIN2007-67581-C02-01).

References

1. E. Amigó, J. Artiles, and J. Gonzalo. Combining Evaluation Metrics with a Unanimous Improvement Ratio and its Application to the Web People Search Cluster-

- ing Task. In *In Proceedings Of The 2nd Web People Search Evaluation Workshop (WePS 2009)*, 2009.
2. J. Artiles, J. Gonzalo, and S. Sekine. WePS 2 Evaluation Campaign: overview of the Web People Search Clustering Task. In *In Proceedings Of The 2nd Web People Search Evaluation Workshop (WePS 2009)*, 2009.
 3. Michael Bloodgood and Chris Callison-Burch. Using mechanical turk to build machine translation evaluation sets. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 208–211, Los Angeles, June 2010. Association for Computational Linguistics.
 4. Balachander Krishnamurthy, Phillipa Gill, and Martin Arlitt. A few chirps about twitter. In *WOSP '08: Proceedings of the first workshop on Online social networks*, pages 19–24, New York, NY, USA, 2008. ACM.
 5. Audrey Le, Jerome Ajot, Mark Przybocki, and Stephanie Strassel. Document image collection using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 45–52, Los Angeles, June 2010. Association for Computational Linguistics.
 6. Management Dell Mit, Chrysanthos Dellarocas, Neveen Farag Awad, and Xiaoquan (michael) Zhang. Exploring the value of online reviews to organizations: Implications for revenue forecasting and planning chrysanthos dellarocas. In *Management Science*, pages 1407–1424, 2003.
 7. Iadh Ounis, Craig Macdonald, and Ian Soboroff. On the trec blog track. In *Proceedings of International Conference on Weblogs and Social Media (ICWSM 2008), Seattle*, 2008.
 8. I. Pollach. Electronic Word of Mouth: A Genre Analysis of Product Reviews on Consumer Opinion Web Sites. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06) Track 3*, 2006.

Test set		Training set	
Entity name	Query	Entity name	Query
Fluent SimplicityAmazon.com	Amazon	alcatel	alcatel
apache	apache	Amadeus IT Group	Amadeus
Apple	Apple	Apollo Hospitals	Apollo
Blizzard Entertainment	Blizzard	armani	armani
camel	camel	barclays	barclays
Canon inc.	canon	BART	BART
Cisco Systems	Cisco	bayer	bayer
CVS/pharmacy	CVS	Blockbuster Inc.	Blockbuster
Denver Nuggets	Denver	Boingo (Wifi for travelers)	Boingo
Deutsche Bank	Deutsche	Bulldog Solutions	bulldog
Emory University	emory	cadillac	cadillac
Ford Motor Company	ford	Craft Magazine	Craft
fox channel	fox	Delta Holding	Delta
friday's	friday's	dunlop	dunlop
Gibson	Gibson	Edmunds.com	Edmunds
General Motors	GM	Elf corporation	elf
Jaguar Cars Ltd.	jaguar	Emperor Entertainment Group	Emperor
John F. Kennedy International Airport	jfk	fender	fender
Johnnie Walker	johnnie	Folio Corporation	folio
kiss band	kiss	Foxtel	Foxtel
Lexus	Lexus	Fujitsu	Fujitsu
Liverpool FC	Liverpool	Harpers	Harpers
Lloyds Banking Group	Lloyd	Impulse (Records)	Impulse
macintosh	mac	lamborghini	lamborghini
McDonald's	McDonald's	linux	linux
McLaren Group	McLaren	Liquid Entertainment	Liquid
Metro supermarket	Metro	Lufthansa	Lufthansa
A.C. Milan	Milan	Luxor Hotel and Casino	Luxor
MTV	MTV	LYNX Express	Lynx
muse band	muse	Mack Group	Mack
Oracle	oracle	Magnum Research	Magnum
Orange	Orange	Mandalay Bay Resort and Casino	Mandalay
Paramount Group	Paramount	Marriott International	Marriott
A.S. Roma	Roma	Marvel comics	Marvel
scorpions	scorpions	mdm (Event agency)	mdm
seat	seat	MEP	MEP
Sharp Corporation	sharp	Mercedes-Benz	Mercedes
sonic.net	sonic	Mercer consulting	Mercer
sony	sony	MGM Grand Hotel and Casino	MGM
Stanford Junior University	stanford	MTA Bike Plus (NYC)	MTA
Starbucks	Starbucks	nikon	nikon
subway	subway	Nordic Airways	nordic
Tesla Motors	tesla	philips	philips
US Airways	US	pierce manufacturing	pierce
Virgin Media	Virgin	Pioner Company	pioneer
Yale University	Yale	Renaissance Technologies	Renaissance
Zoo Entertainment	zoo	Renault	Renault
		Land Rover	Rover
		shin corporation	shin
		Smarter Travel	Smarter
		Southwest Airlines	Southwest
		Yamaha	Yamaha

Table 5. Selected tweets for test and training corpora