ESF - Short Visit Grant - Final Report

Damiano Spina

March, 2012

1 Purpose of the visit

The aim of the visit was to collaborate on the definition, design and development of a corpus of manually annotated microblog posts for entity profiling in Twitter. Given an entity (i.e., a brand, an organization, a person or a product), we want to answer the following question: What do people say about it in Twitter? More precisely, we consider this task as a two-step process: (i) identifying opinion targets on tweets and (ii) linking concepts that are opinion targets to Wikipedia articles. Thus, an opinion-based profiling of an entity in a microblog stream is given by extracting those concepts and entities that occurs as targets of opinions in a stream of microblog posts relevant to the entity.

The applicant is a Ph.D. student at the UNED NLP & IR Group in Madrid, Spain. His MSc thesis, completed in 2011, is focused on company name disambiguation on Twitter [1]. He is currently investigating the task of social media monitoring for on-line reputation management.

The University of Amsterdam has previously worked on fine-grained sentiment annotations and on semantic annotations, both on user generated content [2] and microblogs [3].

2 Description of the work carried out during the visit

The work carried out during the visit can be summarized in three main tasks:

1. **Design and definition of the corpus.** During the visit, we mainly focused on the first step of the entity profiling task, i.e. identifying opinion targets on tweets. The dataset annotated consists of a subset of entities included in the WePS-3 Online Reputation Management dataset [4].

2. **Annotation.** A total of 9,396 tweets related to a subset of 59 entities from the WePS-3 ORM dataset have been annotated at phrase-level. Annotators consider individual tweets related to an entity and manually identify whether the tweet is opinionated and, if so, which part of the tweet is subjective and what the target of the sentiment is, if any.

3. **Short paper submission.** A short paper that describes the developed corpus has been submitted to LREC Language Engineering for Online Reputation Management Workshop[1].

[1] http://www.limosine-project.eu/events/lrec2012

3 Description of the main results obtained

The annotated dataset[2] consists of the tweets of a subset of entities from the WePS-3 dataset, manually annotated at the phrase-level. We aim to identify opinion targets in tweets, related to an aspect of a company. We've defined an *opinion target* as a phrase p that satisfies the following properties:

1. p is an aspect of the entity
2. p is included in a sentence that contains a direct subjective phrase (i.e. an expression that explicitly manifests subjectivity or an opinion)
3. p is the target of the expressed opinion.

A total of 59 companies from the WePS-3 ORM dataset have been annotated, where, for each tweet related to the company, we identify opinion targets and subjective phrases. The resulting corpus includes 9,396 tweets in total, with an average of 159 tweets per company.

3.1 Annotations guidelines

The annotators were asked to indicate the following.

- **Subjectivity:** Tweet-level annotation that indicates whether the tweet contains an explicit opinionated expression.
- **Subjective phrase:** If the tweet is opinionated, identify the phrase that express subjectivity. In our annotation schema, we only considered direct private states [5].
- **Opinion target:** If the tweet contains opinionated phrases, identify the target of the opinion expressed in that phrases.

3.2 Analysis

In total, 9,396 tweets were annotated. Only 1,427 (15.16%) tweets contain subjective phrases and 1,308 (13.82%) contain opinion targets. There are 119 tweets where the annotators identified subjective phrases but not opinion targets. Most of them are tweets containing either emoticons (e.g. `:-)`, `:-(`, `:-/`) or phrases expressing subjectivity at tweet-level (e.g. `LOL`, `Yay!`, `#fail`).

We divide the entities in five groups, based on the number of tweets available for each company (0-10, 11-50, 51-150, 151-300, 301+). For each group C, we count how many companies are part of the group ($|C|$) and the average number of tweets for these entities ($AvgTweets$). Table 1 reports the average of tweets with subjective phrases ($AvgSubj$) and opinion targets ($AvgOpinionTargets$), as well as the averaged percentage of subjective tweets ($Subj\%$) for each group.

We observe that both the average of subjective phrases and opinion targets are directly proportional with the average of relevant tweets. However, the subjectivity ratio tends to stablizes between 10-15% when the number of relevant tweets is higher than 10.

[2]It will be available soon at http://nlp.uned.es/~damiano/datasets/entityProfiling_ORM_Twitter.html

Tweets	$\|C\|$	AvgTweets	AvgSubj	AvgOpinionTargets	Subj (%)
0-10	7	3.57	0.85	0.85	35.11
11-50	11	23.36	3.64	3.09	14.24
51-150	9	96.22	11.77	10.33	11.88
151-300	19	218.68	25.21	23.10	14.22
301+	13	392.54	61.23	56.61	15.80

Table 1: Distribution of subjective phrases and opinion targets, binned by the number of relevant tweets per company.

Future collaborations

So far, the corpus only includes one type of entites: organizations. In future collaborations, we plan to extend the corpus with other types of entities, such as products and people. We also consider to annotated more tweets per entity. We also plan to annotate the corpus linking tweets with Wikipedia pages, following the guidelines used in [3]. This will help us to investigate on the second step of the entity profiling task, i.e., determining salient concepts discussed in a stream of tweets referring to an entity of interest.

Articles resulting from the visit

D. Spina, E. Meij, A. Oghina, B. Minh Thuong, M. Breuss, M. de Rijke. *A Corpus for Entity Profiling in Microblog Posts*, LREC 2012 Workshop on Language Engineering for Online Reputation Management, 2012. (*submitted*)

References

[1] Damiano Spina. Filter keywords and majority class strategies for company name disambiguation in Twitter. Master's thesis, UNED, 2011.

[2] I. Ounis, M. de Rijke, C. Macdonald, G. Mishne, and I. Soboroff. Overview of the TREC-2006 Blog Track. In *Proceeddings of the 15th Text REtrieval Conference (TREC 2006)*, 2006.

[3] Edgar Meij, Wouter Weerkamp, and Maarten de Rijke. Adding semantics to microblog posts. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, New York, NY, USA, 2012. ACM.

[4] E. Amigó, J. Artiles, J. Gonzalo, D. Spina, B. Liu, and A. Corujo. WePS-3 evaluation campaign: Overview of the online reputation management task. In *CLEF 2010 Labs and Workshops Notebook Papers*, 2010.

[5] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2):165–210, 2005.