Quick-and-Clean Extraction of Linked Data Entities from Microblogs

Oluwaseyi Feyisetan
University of Southampton
University Road
Southampton
oof1v13@soton.ac.uk

Elena Simperl
University of Southampton
University Road
Southampton
e.simperl@soton.ac.uk

Ramine Tinati
University of Southampton
University Road
Southampton
r.tinati@soton.ac.uk

Markus Luczak-Roesch
University of Southampton
University Road
Southampton
m.luczak-rosch@soton.ac.uk

Nigel Shadbolt
University of Southampton
University Road
Southampton
nrs@ecs.soton.ac.uk

ABSTRACT

In this paper, we address the problem of finding Named Entities in very large micropost datasets. We propose methods to generate a sample of representative microposts by discovering tweets that are likely to refer to new entities. Our approach is able to significantly speed-up the semantic analysis process by discarding retweets, tweets without pre-identifiable entities, as well similar and redundant tweets, while retaining information content.

We apply the approach on a corpus of 1.4 billion microposts, using the IE services of AlchemyAPI, Calais, and Zemanta to identify more than 700,000 unique entities. For the evaluation we compare runtime and number of entities extracted based on the full and the downscaled version of a micropost set. We are able to demonstrate that for datasets of more than 10 million tweets we can achieve a reduction in size of more than 80% while maintaining up to 60% coverage on unique entities cumulatively discovered by the three IE tools.

We publish the resulting Twitter metadata as Linked Data using SIOC and an extension of the NERD core ontology.

1. INTRODUCTION

The semantic analysis of microblog posts (or 'Making sense of microposts', as a successful workshop series calls it)[1] is one of the most active research topics in the Semantic Web area. With Twitter exceeding all predictions in terms of growth and influence,[2] analysing its vast amounts of user-generated data is essential for anyone aiming to gain a better understanding of how individuals, social groups, governments, and businesses communicate and interact online. However, it is also a challenging task, primarily due to the nature of the content (limited number of characters per post, extreme variation in writing styles, out-of-vocabulary words etc.), and the size and dynamicity of the datasets; all these aspects make the application of off-the-shelf Information Extraction (IE) tools, even when they offer support for semantic technologies or Linked Data, hardly feasible.

Our paper addresses the problem of finding Named Entities in very large micropost datasets (of hundreds of millions, if not billions of items), as a core pre-requisite for more complex social media analytics tasks such as topic, event, and trend detection. Given the sheer scope of the problem, we sought out methods to create a representative sample of posts for faster processing without significant loss in the number of entities retrieved. Such a 'reductionist' approach is important not just for scalability, but also because it gives us an economic means to leverage existing Web-based IE technology and avoid unnecessary HTTP requests. Established online services such as AlchemyAPI,[3] Calais,[4] and Zemanta[5] are known for their IE capabilities, including support for Linked Data URIs on a wide variety of text corpora. However, their use on datasets with millions or even billions of documents often proves impractical due to latencies in the technical infrastructure and service level agreement (SLA) restrictions.

Our work aimed to answer the following research question: How could we carry out large-scale Named Entity Recognition (NER) on microposts by using existing RESTful Web

[1] http://www.scc.lancs.ac.uk/microposts2014/

SEM '14, September 04 - 05 2014, Leipzig, AA, Germany

http://dx.doi.org/10.1145/2660517.2660527

[2] 200 billion tweets per day, referenced by more than 1 million third-party websites, yielding over 30 billion impressions, according to their latest SEC filing. See Twitter Inc, form S-1 at http://www.sec.gov, accessed 2014 − 02 − 17.

[3] http://alchemyapi.com

[4] http://opencalais.com

[5] http://zemanta.com

services without having to analyze each individual post? The basic assumption underlying this line of research was that we could cut down on the number of documents to be analyzed without significant information loss. The very nature of microblogging shows that some posts are either redundant, or do not convey any entity-based information. We exploited this characteristic and devised methods to target and exclude such posts prior to the NER stage in order to improve the overall analysis runtime and optimize on costly API calls. Our approach operates on two fundamental dimensions: (i) the heuristics used to create a representative sample of a given micropost dataset; and (ii) a combination of several NER APIs delivering Linked Data entities. In its current version, the former is centered around eliminating retweets, tweets without proper nouns (as we considered proper nouns to be strong signals for Named Entities), and tweets which we classify as irrelevant or redundant - other criteria could be easily added and tested within the same framework. For the latter we relied on the three IE services mentioned earlier. This choice was informed by the prevalence of these tools in existing literature, including comparative studies such as [17, 18] - again, other APIs or a subset of the three could be equally used and our approach could be configured to take such variations into account.

For the evaluation we compared two basic setups: (i) one which sends a full dataset of tweets to the NER APIs; and (ii) a second one which uses our downscaling method first to compute an information-preserving dataset sample, and then forwards this sample to the APIs. Each of the two setups spans across two evaluation rounds each with five datasets of increasing sizes of $1,000$ to 10 million posts. For each round and setup we analyzed the following metrics: (i) runtime - the time required to process the requests; (ii) downscaling - the number of tweets sent in the scaled down approach; and (iii) extracted entities - the total number of (unique) entities retrieved from each API.

Our results revealed that by cutting the total number of tweets down to 19%, we could still obtain 60% of the entities that would have been returned by processing the entire dataset. The experiments also showed significant runtime improvements for datasets of more than $10,000$ tweets. This was due not only to the reduced size of the problem, but also to the introduction of a computationally efficient implementation of the sampling step, which uses a hashmap index to look up redundant items in constant time, as well as MapReduce distributed processing to identify candidates for elimination at speed.

We released the semantic enrichments as Linked Data using SIOC to capture basic Twitter metadata, and an extension of the NERD core ontology [15] for the types of entities extracted. The dataset is the result of the analysis of 1.4 billion tweets (of which around 573 million are English tweets) and contains $700,555$ unique entities. By publishing this Linked Data set we hope to facilitate the development of future studies of this kind, performed either on the same type of data or on microposts from different social media platforms and possibly using different languages (our approach is currently focusing on English as a first step). In addition, the dataset might prove useful as a scalability benchmark for alignment and interlinking challenges such as the Ontology Alignment and Evaluation Initiative[6] or for the evaluation of IE tools.

It is worthwhile being mentioned that the work proposed in this paper is ongoing. While the general framework and the methods to handle scale are clearly defined, the actual semantic analysis of the Twitter dataset still continues. As noted earlier, we processed around 1.4 billion posts to test and improve our approach. This is part of a much larger corpus of more than 6 billion tweets, which we aim to annotate with Linked Data entities. Independently of this, the preliminary results are very encouraging and show convincing evidence that the approach performs very well on very large datasets, which exceed by orders of magnitude the ones used by previous studies in the literature. As the analysis advances we will make the new findings and the associated Linked Data set available online - the latest updates can be found at `https://sites.google.com/site/twitterentities/`.

The remainder of the paper is structured as follows: Section 2 contextualizes our work with the relevant body of background information and related literature. In Section 3 we describe our generic approach together with the current implementation involving the heuristics and external IE services discussed earlier. Afterwards, in Section 4, we present the experiments undertaken to evaluate the implementation alongside with a description of the research data to allow for reproducing our findings. We conclude in Section 6 with a summary of our main contributions and an outline of future research directions.

2. RELATED WORK

2.1 Information Extraction on Microposts

Previous work in IE for microblogs has adapted traditional Natural Language Processing (NLP) techniques to reflect the specifics of micropost content. This refers mainly to the removal of stop words, retweets, hashtags symbols, ellipses, links, 'user' mentions, as well as out-of-vocabulary words (i.e., 'b4' or 'shuld') [6]. Other common approaches include text tokenizations and optional parts-of-speech (POS) tagging, which use keyword selection to compute the 'link probability' (to Wikipedia article titles) of the tokenized text in order to identify potential entities [16]. Similar methods resort to Wikipedia to match tokenized texts [5], as well as POS tagging to train and identify nouns to be further analyzed [13]. Alternative sources of keyword matching involve Freebase [9], DBpedia [8, 12, 13], and WordNet [16]. The CMU POS Tagger has been developed to handle Twitter-specific vocabulary such as abbreviations (e.g., 'ikr', 'smh'), emoticons (e.g., ':o', ':/'), hashtags, and mentions [2]. Oliviera et al. [3] used five filters (Term, Context, Affix, Dictionary, Capitalization) to decide upon potential entities over continuous Twitter streams. Our approach adopted some of the techniques cited here. However, these studies put more emphasis on IE aspects, while our main aim was to tackle scale while preserving the overall information content. This is also reflected in the design of our analysis framework, which leverages existing IE tools.

2.2 Information Extraction over Big Datasets

[6] `http://oaei.ontologymatching.org/`

Conducting information extraction at scale raises a variety of challenges [1]. Extracting and tagging text corpora of over one terabyte requires hundreds of days on a typical home desktop machine - a shallow syntactic parse on data at such orders of magnitude took 10 machine years according to [14]. While approaches to speed up these processes do exist (for instance, [1] uses indexing and search engines techniques), accurate and timely information extraction Big-Data-style remains difficult to achieve [11, 19].

Out of the studies undertaken in the past, we identified three that share similarities with our research. Dill et al [4] performed 434 million semantic annotations on 264 million Web pages. They ran the documents on a parallel infrastructure with 64 machines at 10,000 documents a second to identify instances that matched a database of 72,000 entities. In an attempt to assess the acceptance of schema.org,[7] [7] embarked on an analysis of a corpus of 733 million Web documents. Whitelaw et al [20] applied supervised learning to Web-scale Named Entity Recognition; they proposed a template-based approach for automatic training data generation, bootstrapped on an initial seed set. The most important difference between these works and ours is in the type of data they consider (Web sites vs. microposts). This fundamentally affects the way documents are processed (e.g., larger word context windows).

One proposal to handle large amounts of Twitter data is presented in [10]. They partitioned tweets and ranked the segments to find potential entities based on specific heuristics. They used Wikipedia articles as frame of reference for entities (thus being less accurate as they operate on a syntactical level), and carried out an evaluation on 4.3 million tweets. Our evaluation uses a corpus of 1.4 billion tweets and samples of varying sizes of up to 10 million posts each (see also Section 4).

2.3 Our Contributions

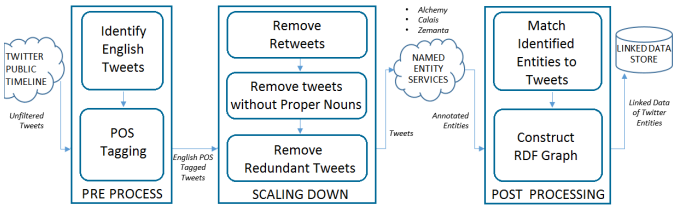
Our work is at the intersection of semantic and large-scale Information Extraction. It makes several important contributions to the state of the art: (i) it proposes a scalable means to process billions of microposts by intelligently exploiting the trade-off between a computationally efficient, but information-preserving sampling method, and rich, but costly Web-based IE services; (ii) it runs, to the best of our knowledge, the largest experiment in semantic enrichment of Twitter data, adding up to 1.4 billion tweets (of which 573 million are English tweets) at the time of the submission, and more than 700,000 unique entities; (iii) it provides a large-scale evaluation of established IE technology on a challenging dataset; and (iv) it creates a Linked Data corpus which can be used for further research in the field, including work on benchmarking data interlinking systems.

3. APPROACH

At a high level, our approach consists of two generic dimensions: (i) a combination of elimination criteria for selecting less informative microposts; and (ii) an aggregation of APIs for information extraction. The first dimension is about creating a representative sample of the original corpus. This is achieved, in the current implementation, in three steps:

[7] https://schema.org/

Figure 1: Overview of Our Approach



(i) remove retweets and mentions, as well as tweets that do not contain proper nouns - our basic assumption is that proper nouns indicate the presence of Named Entities; (ii) discard tweets which contain proper nouns that do not refer to Named Entities (according to a classifier trained on 64,000 entities, see below); and (iii) ignore redundant tweets which do not refer to new entities. The second dimension consists of multiple APIs plugged in for Named Entity Recognition. This workflow is shown in Figure 1, and we describe it in more detail in the remainder of this section.

The workflow is loosely based on the approach by Agichtein [1], which applied search engine techniques to process large amounts of data for information extraction purposes. This is how we applied its main recommendations:

1. *Scan the collection using simplified and efficient rules* - We used simple rules to delete retweets, mentions, # (hashtag symbol), and URLs.

2. *Zoom in on relevant documents and ignore the rest* - We used an iterative approach to locate posts with potential entities, as explained later in Section 3.2.

3. *Use specialized indexes* - We used hashmaps to lookup proper nouns in tweets.

4. *Adopt parallel processing* - We used simple parallel computing techniques to distribute and speed up that part of our analysis pipeline that precedes the NER stage (see also Figure 1).

3.1 Pre-processing the Data

According to Agichtein, the first step involves scanning the data with simple rules. In our case, these rules looked as follows: (i) we filtered out non-English tweets (due to the limitation of 3rd party libraries to deal with other languages, see below); (ii) we stripped out # (hashtag) symbols and URLs that pointed to continuations of tweets. To speed up the process we used a simple MapReduce approach that took advantage of multiple system cores. Data sent to a cluster was processed by each node (on each core).

We then used the CMU POS Tagger[8] to annotate the tweets with parts-of-speech tags. As shown in Table 1 a tweet was tagged without removing the stop words as these (especially prepositions) can serve as an indicator for a word's part-of-speech. Given an original tweet such as *Have you heard about our xmas giveaway at Starbucks on tuesday*, the POS tagger returns a tab separated string. In Table 1, the first

[8] http://www.ark.cs.cmu.edu/TweetNLP/

you	heard	about	our	xmas
O	V	P	D	PN
0.9996	0.9997	0.9924	0.9937	0.7110
giveaway	at	Starbucks	on	tuesday
N	P	PN	P	PN
0.9495	0.9967	0.9723	0.9983	0.9975

Table 1: Parts-of-Speech Tagging Example

row represents the tweet, the second the part-of-speech of each token (e.g., heard: Verb; xmas: Proper Noun), and the third the confidence of the association between the tag and the token.

As noted in [1] POS tagging is one of the most critical performance bottlenecks in large-scale IE. In our case the tagger processed 1,200 tweets per second on an 8 GB RAM machine. This effect was mitigated by running the process in parallel. The choice of POS tagging technology that is tailored for microposts also meant that our approach does not yet support languages other than English. Future releases of the implementation will look into different alternatives for this fundamental pre-processing task, including the GATE POS tagger,[9] which was trained on English tweets, as well as other tagging tools, which we would train ourselves on non-English Twitter data.

3.2 Scaling Down the Data

A central concern in our approach involved deciding which microposts to discard and which ones to keep - this was an essential step to reduce the time required to identify, extract, and match entities. Based upon the results of the POS tagging, we deleted tweets that did not have at least one proper noun, given that a proper noun is a word which, by definition, primarily refers to a Named Entity (e.g., a person, location, organisation).

With all tweets without proper nouns out of the way, we moved on to eliminate all stop words (common words with little novel information value). The tweet *Have you heard about the xmas giveaway at Starbucks on tuesday* was represented as the following list of tuples:

[...(heard, V), (xmas, PN), (giveaway, N), (Starbucks, PN), (tuesday, PN)].

With a further reduced set of tweets, we decided which tweets containing proper nouns to keep for Named Entity Recognition. To achieve this, we developed and trained a simple classifier which was based on a seed set of 3.4 million POS tagged tweets and 64,000 associated entities. The entity set consisted of the raw text, the matched entity, the Linked Data URI, the entity type, and the API that detected the entity.

For each tweet, we extracted the proper nouns (e.g., *(xmas, PN), (Starbucks, PN), (tuesday, PN)* as above). We then checked if any of the proper nouns did not match an entity - in this case, none of the APIs have a Linked Data URI for 'tuesday'. We thus built up a training set of proper nouns that do not seem to refer to Linked Data entities. Any subsequent tweet which had, for example, only 'tuesday' in its set of proper nouns (e.g., *I will see you next tuesday :)*) was discarded.

Our ultimate aim was to find tweets which had novel information content. We understood this property as the presence of an entity that had not been pre-encountered within the same 'context'. The context for an entity (e.g., *xmas* in *Have you heard about the xmas giveaway* was defined as follows: a single word neighbour on both sides (e.g., 'heard xmas giveaway') and a single POS tag on both sides (e.g., 'V xmas N'). We decided to limit the window size to account for the short length of most tweets. The example above, *xmas* in context would be represented as: *('heard xmas giveaway', 'V xmas N')*, where 'V' stands for a verb, and 'N' for a noun.

The approach works as follows: for each new tweet, we extracted the parts-of-speech present to detect the longest continuous proper nouns. For example, given the tweets: *President Barack Obama is on his way to San Francisco* and *President Barack Obama was here on tuesday*, we would obtain the following sets of proper nouns:

[(President Barack Obama, San Francisco); (President Barack Obama)]

Note that the proper noun 'tuesday' would have been discarded in the previous step, as our classifier identified it as non-indicative of Named Entities. The neighbors on either side of the proper noun, and their parts-of-speech would be stored as:

[(President Barack Obama is, President Barack Obama V, to San Francisco)], [(President Barack Obama was, President Barack Obama V)]

While processing the second tweet, we would identify that the only entity 'President Barack Obama', followed by a verb, has been encountered in a previous tweet. Therefore, the second tweet would be marked as redundant and not processed further.

We stored the results of the identified proper nouns and their neighbors in hashmaps; the underlying hash function computes an index using these very proper nouns, which is then used to lookup the existence of such terms in tweets in constant time, hence significantly speeding up the sampling process.

3.3 Information Extraction

The final step of our analysis was outsourced to the 3rd party tools AlchemyAPI, Calais, and Zemanta, which we selected based on their prevalence in the literature studied, including the work by Rizzo et al. [15] on a news corpus, and by Saif et al. [17] on a set of 500 tweets. Due to the limits imposed by the information extraction services, we clustered the tweets in batches to be sent to the respective RESTful APIs. For the evaluation of our approach, these batches consisted of 1,000 tweets selected in the previous step (see Section 4). Table 2 gives an overview of the characteristic dimensions of the three tools when no bespoke service level agreement is negotiated. The use of Zemanta restricted the scope of our

[9] `https://gate.ac.uk/wiki/twitter-postagger.html`

	Alchemy	Calais	Zemanta
Entity Types	324	95	81
Calls per Day	30,000	50,000	10,000
Limit per Call	150 KB	100,000 characters	8 KB

Table 2: Information Extraction APIs

Entity Type	Alchemy	Calais	Zemanta
THING	1,959	-	189,559
ANIMAL	-	-	294
EVENT	130	-	2,925
LOCATION	55,689	16,489	42,434
ORGANISATION	54,795	38,494	51,189
PERSON	212,866	-	86,707
PRODUCT	863	1899	54,690
TIME	120	-	-
TOTAL	326,422	56,882	427,798

Table 3: Number of Entities by Types Retrieved (573,292,411 English tweets)

analysis to English, as this service does not offer support for other languages. Tools such as DBpedia Spotlight were not included in the study following evidence from the literature about their high degree of overlap with AlchemyAPI [18].

We used the NERD ontology [15] to unify the entity type returned by the APIs. NERD consists of a core of 10 classes: *Thing*, *Amount*, *Animal*, *Event*, *Function*, *Location*, *Organisation*, *Person*, *Product*, *Time* and 85 more specialized classes. We augmented the ontology with 82 additional concepts to better match the output of the IE services (21 from AlchemyAPI, 32 from Calais, and 45 from Zemanta). To do so we ran a random set of 5 million tweets through the three APIs, and manually assigned NERD classes to entities marked as `nerd:Unknown`. This was required in 8% of the cases and concerned mostly entity types that have been added to one of the external services after the release of the NERD ontology. Subsequent unknown entities, that is, those which could not be matched to our NERD extension, were marked as `nerd:Thing`.

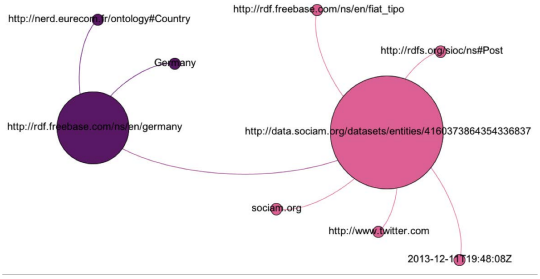
3.4 Publication of the Results as Linked Data

Table 3 gives an overview of the different types, which have been extracted from the 1.4 billion tweets (corresponding to 573 million posts in English) at the time of the submission. The semantic analysis of the corpus resulted in 700,555 unique entities from a total of more than 12 million entities obtained from the APIs. The results show the core entity types aggregated from more fine grain entity types (e.g., the entity type Person, includes sub entities Athletes, Celebrities, Politicians etc).

The results of the analysis will be made available as Linked Data at `http://data.sociam.org`. The raw RDF dump can be downloaded at `http://sociam.org/download/ner_microposts.tar`. Policies of the microblogging platform and its data re-seller did not allow us to publish any original micropost content (including the Status ID) without the user's permission or in conflict with display requirements. Consequently, we only released the extracted entities and not the original tweets. While we are exploring ways to enrich the dataset with at least parts of the primary corpus while complying to the licensing agreements of the Twitter data provider, we believe the published metadata already presents a valuable resource for further studies in this and related fields, including more accurate, concept-based topic analysis and trend detection.

The dataset looks as follows: each micropost is associated to a resource of type `sioc:Post` identified by a URI in the `data.sociam.org` namespace, which is owned by our research group. This resource is associated to: (i) a timestamp when the original micropost was created via `dc:created`; (ii) the platform on which the post was originally published via `sioc:has_container`; and (iii) the entities extracted, which are identified by external resource identifiers served by DBpedia, MusicBrainz, Freebase, and OpenCalais via `sioc:topic`. In addition, entities are typed conforming to classes within the NERD ontology. An example of a micropost instance as RDF is shown in Figure 2.

Figure 2: Example RDF Graph



4. EXPERIMENTS AND EVALUATION

We developed an approach, which allows for leveraging multiple RESTful APIs to extract Linked Data entities from microblog posts. We were primarily concerned with scalability, aiming to optimize the costs of the computation, be those technical costs such as processing time; or real costs depending on the SLAs of the external APIs. This was achieved with the help of a method to eliminate redundant or useless microblog posts. We ran two rounds of experiments on datasets of several sizes to investigate the following aspects:

Runtime: How does the reduction of posts affect the runtime of the system per API and overall?

Downscaling: How does the proportion of eliminated posts develop depending on the size of the processed data sample?

Extracted entities: How does the absolute number of extracted entities (unique and overlapping) develop depending on the size of the processed data sample?

By answering these questions for our system we gained insight into the influence of its two core design dimensions (sampling heuristics and NER services) on the overall performance. The results will also inform future work on alternative configurations of the framework, most importantly other elimination criteria, as well as different ways to cluster tweets to be processed by the external APIs. As noted earlier, due to limitations in the document size to be processed

by each API (see Table 2), per API call we used documents containing a concatenation of 1,000 chronologically ordered tweets.

It is worthwhile mentioning that we did not use metrics such as precision, recall and F-measure in the evaluation. The reason for this experimental design choice is that we rely on external NER services, which have been subject to several comparative studies in the past [15, 17, 18].

4.1 Research Data

The dataset used in this study is a collection of tweets extracted from a double 'garden hose' (20%) stream of the full Twitter data stream. This represents a random sample pre-computed by Twitter and purchased from a popular reseller. The collection of the data was captured in real-time and stored at the point of retrieval. Each record within the dataset contains a unique tweet identifier, the tweet user, the body (message) of the tweet, and the detected language ('en', 'fr', etc.) of the tweet. Best practises were employed to ensure the data collection was performed reliably, thus providing a consistent and complete pre-processed dataset for further analysis. We then extracted a total of 1,433,230,956 tweets between the time period of the 29th November 2013 to the 9th February 2014. We chose this time slice as it featured a consistent volume of daily tweets, with minimum gaps in collection (less than 5 days in total where no tweets were gathered). 573,292,411 tweets were identified as English (pre-classified by Twitter). We recognise that language identification on microposts is an issue in itself. However, our experience with Twitter's language metatag, especially in identifying English tweets, yielded few false positives. It appears Twitter favours language identification precision over recall - thus assigning a metatag of "U" to unknown languages.

4.2 Evaluation Method

From the 573 million English tweets in the research data corpus we created five evaluation datasets with varying scale (1,000 to 10 million tweets) in order to gain a better understanding of the behaviour of the different components of our approach in relationship to the number of posts to be processed. We understand the potential of bias in our sample set selections (e.g., November posts contained several references to Thanksgiving and December to Xmas). To mitigate this, we ran our evaluation twice (over different datasets with 1,000 to 10 million tweets), using various sections of the overall corpus. Each of the two rounds of experiments followed two setups for each of the five evaluation datasets: (i) in the first setup we sent all tweets to the IE tools (AlchemyAPI, Calais, Zemanta); (ii) in the second setup we only used the downscaled version of each dataset.

The results discussed in the remainder of the section refer to one of the two rounds of experiments. The second round was designed to confirm the findings of the first; the differences in runtime and number of entities were insignificant and did not add to the evaluation.

4.3 Results

Table 4 below shows some highlights of the pre-processing step. An average of 40% of Twitter seems to be using English as a language. In addition, 39% of these English tweets contain at least one proper noun, according to the CMU POS tagger.

Number of English Tweets	Tweets with Proper Nouns
1,000	393 (39.3%)
10,000	3,936 (39.3%)
100,000	39,845 (39.8%)
1,000,000	389,304 (38.9%)
10,000,000	3,703,644 (37%)

Table 4: English Tweets with at Least One Proper Noun

Table 5 shows the results of the NER APIs usage. Each subsection of the table records the following outcomes for each of the five evaluation datasets.

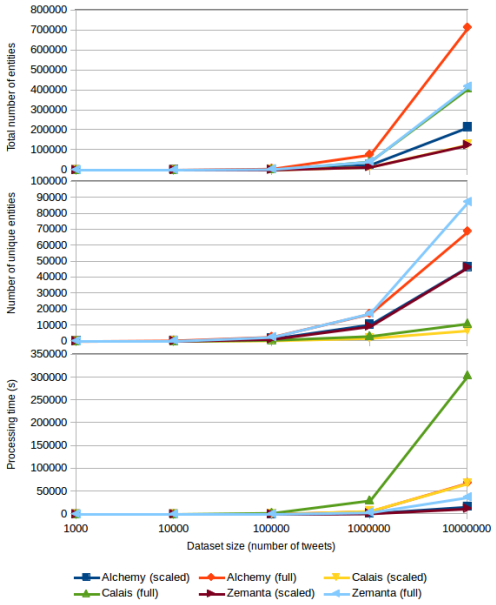
1. Runtime - The time (in seconds) required to receive the results from each service individually (column 'Runtime'), and all three in parallel (row 'Processing time').

2. Downscaling - The number of tweets sent in the scaled down approach for each dataset.

3. Extracted entities - Total and unique entities extracted from each API. We used a combination of the matched entity text and the NERD classes to determine unique entities (e.g., `brandy` typed as `nerd:Person` is different from `brandy` typed `nerd:Thing`).

Figure 3 depicts the course of the most important quantitative criteria mentioned in Table 5 for each of the three APIs.

Figure 3: Performance of the NER APIs (x: number of tweets; y: number of entities delivered)



5. DISCUSSION

In this section we discuss the most important findings of our evaluation and its implications for future research.

	Scaled Down Set			Full Dataset		
API	**Total**	**Unique**	**Time**	**Total**	**Unique**	**Time**
1,000 tweets						
ALC	64	64	42.61	32	71	6.51
CAL	32	32	51.49	33	33	25.07
ZEM	78	78	46.79	30	30	3.83
Tweet	276 (27.6%)			-		
Time	52.94			27.38		
Entity	2			1		
10,000 tweets						
ALC	377	327	63.41	670	499	64.29
CAL	162	128	131.71	333	201	285.15
ZEM	217	210	59.53	413	354	40.82
Tweet	2,443 (24.4%)			-		
Time	122.03			285.33		
Entity	13			17		
100,000 tweets						
ALC	2,839	1,701	233.62	6,605	2,788	647.93
CAL	1,405	490	746.83	3,562	810	2,971.12
ZEM	1,659	1,351	166.40	4,014	2,625	388.28
Tweet	20,112 (20.1%)			-		
Time	657.96			2,986.61		
Entity	85			130		
1,000,000 tweets						
ALC	26,443	10,497	2,302.58	76,776	17,226	6,770.44
CAL	15,516	1,983	7,217.51	42,800	3,265	30,438.74
ZEM	15,344	9,385	1,723.82	41,056	17,459	3,952.98
Tweet	194,326 (19.4%)			-		
Time	6,843.57			30,438.74		
Entity	470			632		
10,000,000 tweets						
ALC	214,387	46,610	16,636.41	918,244	123,475	74,401.43
CAL	127,648	6,770	72,677.66	546,080	16,594	330,342.38
ZEM	124,941	46,321	12,336.75	541,744	128,665	43,069.41
Tweet	1,309,170 (13.1%)			-		
Time	12,336.75			330,342.38		
Entity	1,496			3,861		

Table 5: Analysis for Data Samples $1,000$ to $10,000,000$ Tweets

5.1 Runtime

As it can be seen in the results obtained for the $1,000$ tweets dataset in Table 5, our approach does not pay off for very small datasets. From $10,000$ tweets onwards we see significant runtime gains due to an efficient implementation of the sampling step paired with the reduced size of the dataset that is subject to the IE step. Our experiments also revealed differences in processing time, which need to be taken into account if the outcomes of multiple tools are to be reconciliated into a unique set of entities. Zemanta offered the fastest processing time, with AlchemyAPI coming in at a close second. The Calais API was orders of magnitudes slower. It would be interesting to expand the scope of the evaluation with a more detailed study into the types of entities that could be most effectively obtained from each API. This would require a more thorough understanding of the inner workings of each service together with a more refined method for the analysis of redundant tweets. As the overlap between the three tools is very low, an additional option would be to offer more liberal means to identify similar entities, which would then allow us to use each API for a designated purposes.

5.2 Downscaling

We could observe a downward trend in the number of tweets that are sent to the APIs after the elimination process. The percentage went down steadily from 27.6% in the smallest dataset, to 13.1% in the 10 million dataset. Ongoing experiments on even larger datasets of up to 80 million tweets processed at once led to reductions of more than 89%. This confirms our basic research hypothesis that scaling down the scope of the analysis is feasible and does not result in a proportional loss in the number of entities retrieved. Furthermore, as shown in Table 4, the share of tweets containing proper nouns remains fairly constant. As such, the steady decrease in the number of processed microposts can be safely attributed to a larger degree of redundancy in the number of new Named Entities. At this stage our approach offers only a rough, though indicative quantitative characterization of the information retained by our sampling method. It would be interesting to be able to learn more about the types of entities that are potentially lost via the discarded tweets, and the 'importance' of such entities (expressed in information-theoretical terms of alike).

5.3 Entity Extraction

The number of unique entities discovered decreased as the dataset size and the number of discarded tweets increased. In the 10 million tweets setting, 13.1% of the original dataset still contained 40% of the entities that were mentioned in the full dataset. Downscaling the one million tweet set to 19.4% of its scope resulted in 60% of the total entities. The Linked Data entities were returned by the APIs together with their disambiguated types. As noted in previous studies [18] the degree of overlap between the three tools is very low, as each of them resort to different knowledge bases for type assignment (AlchemyAPI uses DBpedia, Zemanta Freebase, and Calais their own OpenCalais vocabulary).

6. CONCLUSIONS AND FUTURE WORK

In this paper, we addressed the problem of large-scale Named Entity Recognition on microblog content. We introduced techniques for downscaling a corpus of microposts by discarding redundant or irrelevant elements. We used a configurable set of external Information Extraction APIs to identify Named Entities and their corresponding Linked Data URIs and types in microposts. We evaluated runtime, information loss via sampling, and number of (unique) entities in two rounds of experiments on datasets of varying sizes with encouraging results. In a nutshell, the experiments clearly confirmed the feasibility of our idea. They showed that the heuristics we applied do create meaningful samples of Twitter data. For one million tweets a sample of around 19% of the original dataset offered a 60% coverage in Linked Data entities. For datasets an order of magnitude bigger a chunk of 13.1% of the tweets still referred to 40% of all entities. An efficient implementation of the downscaling step paired with the reduced scope of the IE analysis led to significant performance gains, which increase with the size of the dataset (4.45 times faster for one million tweets, and a factor of 26.78 for 10 million tweets). As discussed earlier, immediate future work will include additional experiments to study in more detail the features of each APIs and a more refined

way to use them in combination, alongside the evaluation of alternative clustering methods (for grouping tweets into more coherent documents to be sent to the NER APIs) and support for languages other than English. The most up-to-date results of our semantic analysis will be made available at `https://sites.google.com/site/twitterentities/`.

7. REFERENCES

[1] Eugene Agichtein. Scaling information extraction to large document collections. *IEEE Data Eng. Bull*, 28:3–10, 2005.

[2] Amitava Das, Utsab Burman, Balamurali Ar, and Sivaji Bandyopadhyay. NER from Tweets: SRI-JU System. In *Proceedings of the Concept Extraction Challenge at the Workshop on 'Making Sense of Microposts'*, page 62, 2013.

[3] Diego Marinho de Oliveira, Alberto H.F. Laender, Adriano Veloso, and Altigran S. da Silva. FS-NER: A Lightweight Filter-stream Approach to Named Entity Recognition on Twitter Data. In *Proceedings of the 22nd International Conference on World Wide Web Companion*, WWW '13 Companion, pages 597–604, 2013.

[4] Stephen Dill, Nadav Eiron, David Gibson, Daniel Gruhl, R. Guha, Anant Jhingran, Tapas Kanungo, Sridhar Rajagopalan, Andrew Tomkins, John A. Tomlin, and Jason Y. Zien. SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation. In *Proceedings of the 12th International Conference on World Wide Web*, pages 178–186. ACM, 2003.

[5] Yegin Genc, Winter A. Mason, and Jeffrey V. Nickerson. Classifying Short Messages using Collaborative Knowledge Bases: Reading Wikipedia to Understand Twitter. In *Proceedings of the Concept Extraction Challenge at the Workshop on 'Making Sense of Microposts'*, pages 50–53, 2013.

[6] Bo Han and Timothy Baldwin. Lexical Normalisation of Short Text Messages: Makn Sens a #Twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, HLT '11, pages 368–378. ACL, 2011.

[7] Silviu Homoceanu, Felix Geilert, Christian Pek, and Wolf-Tilo Balke. Any Suggestions? Active Schema Support for Structuring Web Information. In *Database Systems for Advanced Applications*, pages 251–265. Springer, 2014.

[8] Amir Hossein Jadidinejad. Unsupervised Information Extraction using BabelNet and DBpedia. In *Proceedings of the Concept Extraction Challenge at the Workshop on 'Making Sense of Microposts'*, pages 54–56, 2013.

[9] David Laniado and Peter Mika. Making Sense of Twitter. In *Proceedings of the 9th International Semantic Web Conference*, pages 470–485. Springer, 2010.

[10] Chenliang Li, Jianshu Weng, Qi He, Yuxia Yao, Anwitaman Datta, Aixin Sun, and Bu-Sung Lee. TwiNER: Named Entity Recognition in Targeted Twitter Stream. In *Proceedings of the 35th Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 721–730. ACM, 2012.

[11] Songyu Ma, Quan Shi, and Lu Xu. The Research of Web Parallel Information Extraction Based on Hadoop. In *Proceedings of International Conference on Computer Science and Information Technology*, pages 341–348. Springer, 2014.

[12] Pablo N. Mendes, Dirk Weissenborn, and Chris Hokamp. DBpedia Spotlight at the MSM2013 Challenge. In *Proceedings of the Concept Extraction Challenge at the Workshop on 'Making Sense of Microposts'*, pages 57–61, 2013.

[13] Óscar Muñoz-García, Andrés García-Silva, and Óscar Corcho. Towards Concept Identification using a Knowledge-Intensive Approach. In *Proceedings of the Concept Extraction Challenge at the Workshop on 'Making Sense of Microposts'*, pages 45–49, 2013.

[14] Deepak Ravichandran. *Terascale Knowledge Acquisition*. PhD thesis, Los Angeles, CA, USA, 2005. AAI3196880.

[15] Giuseppe Rizzo and Raphaël Troncy. NERD: A Framework for Unifying Named Entity Recognition and Disambiguation Extraction Tools. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 73–76. ACL, 2012.

[16] Sandhya Sachidanandan, Prathyush Sambaturu, and Kamalakar Karlapalem. NERTUW: Named Entity Recognition on Tweets using Wikipedia. In *Proceedings of the Concept Extraction Challenge at the Workshop on 'Making Sense of Microposts'*, pages 67–70, 2013.

[17] Hassan Saif, Yulan He, and Harith Alani. Semantic Sentiment Analysis of Twitter. In *Proceedings of the 11th International Conference on The Semantic Web*, pages 508–524. Springer, 2012.

[18] Seth van Hooland, Max De Wilde, Ruben Verborgh, Thomas Steiner, and Rik Van de Walle. Exploring entity recognition and disambiguation for cultural heritage collections. *Literary and Linguistic Computing*, 2013.

[19] Henning Wachsmuth, Benno Stein, and Gregor Engels. Constructing Efficient Information Extraction Pipelines. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 2237–2240. ACM, 2011.

[20] Casey Whitelaw, Alex Kehlenbeck, Nemanja Petrovic, and Lyle Ungar. Web-scale Named Entity Recognition. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 123–132. ACM, 2008.