Entity Grouping for Accessing Social Streams via Word Clouds

Martin Leginus[1], Leon Derczynski[2], and Peter Dolog[1]

[1] Department of Computer Science, Aalborg University, Selma Lagerlofs Vej 300, 9200 Aalborg, Denmark, {mleginus,dolog}@cs.aau.dk,
[2] Department of Computer Science, University of Sheffield, S1 4DP, United Kingdom, leon@dcs.shef.ac.uk

Abstract. Word clouds have been proven as an effective tool for information access in different domains. As social media is a main driver of large increase in available user generated content, means for accessing information in such content are needed. We study word clouds as a means for information access in social media. Currently-used clouds that are generated from social media data include redundant and mis-ranked entries, harming their utility. We propose a method for generating improved word clouds over social streams. In this method, named entities are detected, disambiguated and aggregated into clusters, which in turn inform cloud construction. We show that word clouds using named entity clusters attain broader coverage and decreased content duplication. Further, an extrinsic evaluation shows improved access to data, with word clouds having grouped named entities being rated more relevant and diverse. Additionally we find word clouds with higher Mean Average Precision (MAP) tend to be more relevant to underlying concepts. Critically, this supports MAP as a tool for predicting cloud quality without needing a human.

Key words: word clouds, recognized named entities, user evaluation, social media, social stream access

1 Introduction

A word cloud is a visual information retrieval interface which presents prominent and interesting terms from the underlying data collection. Word clouds allow quick access and exploration over document collections [1] and reduce information overload [2]. There are various studies about tag cloud generation from folksonomy data [3, 4], but few studies available about word clouds generated from user generated content on social media [5].

To investigate information access over social media, we investigate the "model organism" of this data type, Twitter [6], a worldwide popular online social network where users publish daily an enormous amount of content (upwards of 600 million pieces of content per day). Therefore, Twitter users often face information overload while searching, browsing and exploring tweets [7]. To enhance information access to relevant tweets, one might leverage word-cloud based retrieval interfaces. Word clouds can be intuitively employed for browsing of underlying collection of tweets and at the

same time enabling access only to relevant content. For instance, the interactive browsing interface Eddi, where a word cloud is a core component of the interface [7], helps to decrease information overload. According to users, Eddi gives a more efficient and enjoyable mode of browsing the enormous amount of user stream tweets. To further improve usefulness of word clouds, a personalized cloud generation is proposed [5]. The suggested framework combines different user past actions (user past tweets and retweets) with negative user past preferences (tweets read but not indicated as relevant) to generate personalized word clouds. Such personalized clouds enhance access to relevant tweets when compared with state-of-the-art non-personalized approaches. Authors find user past retweets as more useful for personalization of clouds than user past published tweets. In addition, negative past preferences when combined with user past positive actions further improve the quality of word clouds. Similarly, word clouds ease e-health monitoring when browsing large collections of tweets [8].

Despite the benefits of word clouds for accessing and browsing social stream data, it remains a difficult type of text to handle. As a result of the diversity of language choice and spelling present in social media, end users are often presented with several different terms that refer to the same entity or concept, each term using different syntax and form; this leads to an increase lexical sparsity for the same degree of conceptual sparsity [9].

Hence, conventional methods of generating word clouds lead to undesirable results when applied to social stream text. In particular, variations of proper nouns create duplicated clusters, each of reduced prominence. Compounding the issue, variety in expression is increased by tight space constraints in some formats (like Twitter's 140-character limit) and by social media's generally informal, uncurated setting, as well as the inclusion of quasi-word hashtags [10, 11]. For example, the football club *"Manchester United"* may also be referred to as *"MUFC"* and *"Man U"*. Adding entries for each of these leads to a decrease in the prominence of this key concept, while also taking up space in the cloud and thus reducing its eventual diversity.

Redundancies in the word cloud might lead to user confusion and an inability to effectively browse, explore and retrieve other relevant content. Therefore, two aspects should be considered when designing word cloud generation algorithms. Primarily, one aims to condense divergent terms describing the same concept into a single term. Also, a cloud's high-level diversity must be maximised, so the cloud gives a broad account of topics in the collection. These two conflicting requirements must be balanced to achieve an optimal word cloud.

The aim of this study is to improve word cloud generation by grouping co-referent entity expressions across multiple documents, applying existing named entity recognition systems in a novel fashion and grounding terms in this difficult genre to linked data resources. We systematically study the benefits of grouped named entities on word cloud generation, and investigate the role of unsupervised hierarchical clustering in finding candidate entity synonyms. This work builds on a previous conference paper version, [12]. We use three established synthetic metrics – Coverage, Overlap and Mean Average Precision (MAP) – for word clouds generated from social media data [4, 5]. Further, to verify the findings of the synthetic evaluation, we perform a user study com-

paring clouds with grouped and un-grouped named entities. The main contributions and findings of this paper are:

- The best performing DivRankTermsEntities method significantly increases Coverage with respect to the baseline method ($p = 0.0363$) and significantly decreases Overlap ($p = 0.000094$) (decreased redundancies). In addition, access to relevant documents is improved.
- Word clouds with grouped named entities are significantly more relevant ($p = 0.00062$, one sample t-test) and diverse ($p = 0.003$, one sample t-test).
- Users report that word clouds with grouped named entities that attain higher levels of MAP are more relevant than the word clouds with decreased levels of MAP. Hence, the MAP metric should be considered and measured when designing word cloud generation methods.

The structure of the paper is as follows. Section 2 provides a brief description of relevant related work. Section 3 describes a general process of word cloud generation and points out the focus of our work, and presents a method for word cloud generation with grouped entities, experimenting with both commerical and unsupervised entity alias extraction. In Section 4, we describe graph-based word cloud generation which is the underlying framework for the later evaluation. Section 5 presents the findings from an offline evaluation of generated word clouds from TREC2011 microblog collection as well as results of the performed user study. Finally, we discuss the paper's contributions as well as possible limitations of this work in Sections 6 and 7.

2 Related work

2.1 Word cloud generation

Tag cloud generation from folksonomy data has been thoroughly researched. Several tag cloud generation methods are proposed [3, 4] and even synthetic metrics expressing tag cloud quality designed [4]. There are few studies that explore the benefits of word clouds for browsing social stream data. For instance, the browsing tool Eddi, where word cloud is a core component of the interface [7], helps to decrease information overwhelm. Similarly, word clouds are useful for the detection of epidemics when browsing thousands of tweets is needed [8].

Crowdsourcing has been used to recognize named entities in tweets [13]. This study reports that word clouds with named entities recognized by human workers are considered better. This supports our motive to promote and improve the handling of named entities in word clouds recognized in tweets.

Our contribution beyond the study from [13] is threefold. First, we perform grouping of recognized named entities, which has a positive impact on the generated word clouds. Second, we systematically study how to generate word clouds with named entities and measure performance using multiple metrics. Third, we compare these measured performances with user ratings, to discover relations between metrics and the user's perspective.

2.2 Social stream text repair

Social stream text is noisy, and difficult to process with typical language processing tools [10]. Consolidation of the varying expressions used to mention entities is possible, over large well-formed corpora [14]. Achieving this over social streams presents new challenges, in terms of the reduced context and heightened diversity of expression [15, 16, 17]. The field continues to be active, with the ACL 2015 W-NUT shared task being dedicated to text normalisation and attracting many submissions [18] – the other task at this venue being named entity recognition in social streams, which we also focus on in this work.

We propose a simple consolidation technique and explore its positive impact on the word cloud generation. Other potential methods we could employ to improve cloud quality are normalization and co-reference. Normalization [19] applies to many low-frequency terms, and as a result has a low impact with named entities. Also, while normalisation can compare minor spelling mistakes, it typically does not condense highly orthographically different expressions of the same entity. Co-reference requires context to operate – something that is absent in short social media stream messages. Mapping keywords to unambiguous entity references is difficult, but understood [20].

3 Word clouds generation with grouped named entities

The process of generating word clouds from social media data is comprised of several subsequent steps:

1. *Data collection* where underlying documents are aggregated with respect to a user query, profile or trending topic. Often, the whole document collection might be used for a word cloud generation. In this work, we aggregate tweets for word cloud generation with respect to a user query.
2. *Data preprocessing* where extracted terms or phrases can be clustered, lemmatised or normalised. Documents can be further enriched or annotated with recognized named entities. The aim of this work is to investigate how recognized named entities detected during this phase impact the following word cloud generation.
3. *Word cloud generation* where the most relevant and important terms from the underlying collection are selected and consequently a word cloud is generated. Different word selection methods can be applied [4], [3], [5].

The goal is to explore how recognized and grouped named entities from the *Data preprocessing* phase affect consequent word cloud generation. Do grouped named entities improve the quality of word clouds in terms of Coverage, Overlap and enhanced access to relevant tweets? Which word cloud generation method gives best results when using named entities? We transform these research questions into the following two hypotheses:

– *H1: Word clouds with grouped recognized named entities improve Coverage, Overlap and Mean Average Precision of generated word clouds.*

– *H2: Word clouds with grouped recognized named entities are more relevant and more diverse with respect to a provided query from the user perspective.*

In the following, we describe a method for grouping recognized named entities from tweets.

3.1 Grouping named entities

Conventional named entity recognition is not sufficient due to the nature of Twitter data [21]. Standard named entity recognition approaches do not perform well on tweets because of the error prone structure (misspellings, missing capitalization or grammar mistakes) and their short length. We propose a method that aims to recognize named entities, to link the possible aliases and consequently to generate a word cloud with the recognized and linked named entities. This method can be thought of as a *Data preprocessing* step when generating word clouds over data from social streams.

We combine standard named entity recognition tools with linked data. Alternative names for recognized entities are exploited for term cluster creation for each named entity. A canonical term from an entity term cluster is selected and, if relevant and prominent enough, it is presented in the final word cloud. The method is summarized as follows:

1. Gather a tweet collection – a set of tweets corresponding to a certain trending topic or a query on Twitter.
2. Recognise named entities (**NER**) and disambiguate them (**entity linking**) using the TextRazor service, which performs this task relatively well [21].[1]
3. Using linked data, find alternative names for the recognised entity. We used Freebase's [22] `aliases` field for this. For instance, for the entity *Manchester United FC* the following aliases might be retrieved *Man United*, *Manchester United*, *Man Utd*, *MUFC*, *Red Devils*, *The Reds* or *United*.
4. Perform lemmatisation to group together all the inflicted forms of a word to exploit only the base form of the term.
5. Using the aliases, build a term cluster for each entity, containing e.g. *Manchester United*, *Man U*, *MUFC*.
6. Find canonical names, such as *Manchester United FC*.
7. Generate the "condensed" cloud with aggregated counts of entity mentioned frequencies with some word cloud generation technique.

This may be performed as a general-purpose technique, and also to "targeted" streams, e.g. where tweets are filtered based on user-defined criteria such as keywords or spatial regions.

3.2 Distributional named entity grouping

To examine the viability of alternatives to the commercial TextRazor, we also investigate whether unsupervised clusterings can provide entity groups. We run a large corpus

[1] See www.textrazor.com

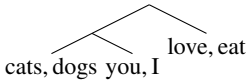

Fig. 1: A binary, hierarchical clustering of semantically similar entries. Each leaf corresponds to a cluster of words (i.e., a "class") and leaves near to their common ancestors correspond to clusters that are similar to each other.

of Twitter data sampled from the twitter "garden house" [23] through Brown clustering [24]. This technique captures distributionality through mutual information and uses this as a metric for agglomerative bottom-up hard clusterings of terms found in a corpus. The end result is a binary tree, with terms as leaves, and subtrees holding semantically similar properties (Figure 1). Paths from the root are described using a bitstring, which has the virtue of being able to estimate semantic similarity of two paths be the length of their common prefix. The transition from distributional to semantic similarity comes from the feature of language that the meaning words can be determined from the words around them [25, 26].

We built a Brown clustering with 2500 classes, an optimal value in many situations [27], from approximately 10^9 English language tweets sampled within 2009-2014. We used langid.py [28] for language filtering. From this hierarchical clustering, we selected candidate entity terms and qualitatively explored how well entity surface forms were grouped by this technique.

Note that increased precision in clustering comes at potentially large computational cost, and so the accuracy of our data is constrained by the amount of time available; Twitter is a challenging environment due to proliferation of word types (i.e., surface forms) in relation to newswire. Additionally, the clustering is for single word types, and so multi-word expressions are out (e.g. *red devils*).

We noted that different expressions of the same entity were represented closely in the resulting tree.[2]

```
11111110110     MUFC    2147
11111110010     #mufc   3131
11111110110     #MUFC   5470

100010010100    mufc    114
```

In this case, *MUFC*, *#mufc* and *#MUFC* are all close to each other, while *mufc* belongs to another cluster. Interestingly, the *mufc* cluster is rich in football team-related terms; it contained 4800 items, the most frequent of which looked like this:

```
100010010100    Milan     5148
100010010100    Rangers   5839
100010010100    Barcelona 12869
100010010100    Chelsea   17589
100010010100    Arsenal   18603
100010010100    United    34614
100010010100    Liverpool 23231
```

[2] In these output excerpts, the columns are: bitstring; word; frequency in dataset.

The technique is consistently effective at grouping alternative spellings of the same terms, which is very useful in the context of social media. this opens up further opportunities for consolidating the noise intrinsic to social media – frequently used words will grow into clusters containing the majority of their spelling variants. Thus, not only do Brown clusters offer means of disambiguating entity co-references for consolidating terms in tag clouds; it is also possible to consolidate non-named-entity terms, both with regard to their spelling and also their semantics.

```
0111100000  tm         2414
0111100000  tomorow    2420
0111100000  tommorow   7526
0111100000  tmr        8387
0111100000  tmrw       12075
0111100000  tomorrow   571411

0111111110      mexico       1
0111111111110   Mexicooooo   3
0111111111110   mexicooo     12
0111111111110   mexicoooo    15
```

While – de facto – Textrazor provides useful groundings and Freebase provides useful aliases, we propose use of Brown clustering as a potential source of alternative term generation and capture.

4 Graph-based word cloud generation

In this section, we describe a graph-based method for generating word clouds with and without entity grouping. The benefits of graph-based word cloud generation are following. First, the method identifies relevant and important keywords in underlying text collections. Several studies have empirically demonstrated the benefits of graph based methods over standard popularity or TF-IDF word cloud generation methods [3, 5, 29]. Second, graph-based methods allow biasing of word cloud generation toward user preferences or search queries. Our graph-based selection methods firstly transforms terms space into a graph. Then, the stochastic ranking of vertices in the graph is performed. In this work, we consider only global ranking but the proposed methods can be easily applied to biased graph-based ranking e.g., biased towards a user query or a user profile.

4.1 Graph-based creation

Extracted terms from underlying tweets are used to build a graph where each term is a graph vertex. If two terms (vertices) co-occur at least α times, we consider these two terms as similar. Eventually, for each similar term pair, two directed edges are generated $t_1 \rightarrow t_2$ and $t_2 \rightarrow t_1$. Hence, edges capture co-occurrence relations between individual terms.

4.2 Graph-based ranking

Graph-based ranking of terms simulates a stochastic process i.e., random traversal of the terms in the graph. We use a PageRank-style algorithm [3], but any other algorithm based on random traversal of the graph could be employed. The aim is to estimate the global importance of a term t. If needed, it is possible to bias ranking towards user preferences through a vector of prior probabilities $\mathbf{p_p}$. For global graph-based ranking, i.e. without introduced bias, we set each entry in $\mathbf{p_p} = \{p_1 \dots p_{|V|}\}$ to $\frac{1}{|V|}$ where V is the set of all graph vertices. The sum of prior probabilities in $\mathbf{p_p}$ is 1. A random restart of stochastic traversal of the graph is assured with a back probability β which determines how often a random traversal restarts and jumps back to a randomly selected (following $\mathbf{p_p}$ probability distribution) vertex in the graph. So, the β parameter allows adjustment of bias toward user preferences or to vertices that are globally relevant in the underlying graph. To simulate random traversal of the graph, iterative stationary probability is defined as:

$$\pi(v)^{(i+1)} = (1-\beta)\left(\sum_{u=1}^{d_{in}(v)} p(v|u)\pi^{(i)}(u)\right) + \beta\mathbf{p_p} \tag{1}$$

where $\pi(v)^{(i+1)}$ is a probability of visiting node v at time $i+1$, $d_{in}(v)$ is the set of all incoming edges to node v and $p(v|u)$ is a transition probability of jumping from node u to node v. In this work, a transition probability is set to $p(v|u) = \frac{1}{d_{out(u)}}$ for nodes v that have an ingoing edge from node u, otherwise $p(v|u)$ equals 0. The resulting global rank of a term t after convergence is considered as relevance of t i.e.;

$$I(t) = \pi(t) \tag{2}$$

Top-k ranked terms are then used for word cloud generation where the ranking score indicates the prominence of the term in a word cloud.

5 Evaluation

We retrieved available tweets with relevance judgments from TREC2011 microblog collection [30] during August 2014. Although some tweets were not available during retrieval, we compare results over the same corpus. We do not consider the missing tweets as a limitation of our evaluation – see [31]. The relevance judgments for TREC2011 microblog collection were built using a standard pooling technique. For TREC the relevance of a tweet with respect to a query was assessed with a three-point scale; 0: irrelevant tweet, 1: relevant tweet and 2: highly relevant tweet. In this work, we consider both relevant and highly relevant tweets as equally relevant.

5.1 Metrics

We evaluate individual aspects of generated word clouds using the synthetic metrics introduced in [4, 5]. The generated word cloud with k terms is denoted as WC_k. A term

t links to a set of tweets Tw_t. Tw_{t_q} is the set of all tweets that are associated with a query phrase t_q.

Similarly, $TwREL_{t_q}$ is the set of all relevant tweets for the query t_q.

The first metric is *Coverage*, defined as:

$$\text{Coverage}(WC_k) = \frac{|\cup_{t \in WC_k} Tw_t|}{|Tw_{t_q}|}, \tag{3}$$

where the numerator of the fraction is the size of the union set. The union set consists of tweets associated with each term t from the word cloud WC_k. $|Tw_{t_q}|$ is the number of all tweets that are associated with a query phrase t_q. The metric ranges between 0 and 1. When a Coverage for a particular word cloud WC_k is close to 1, the majority of tweets are "covered" i.e., linked from the word cloud WC_k.

Overlap of WC_k: Different words in WC_k may be linking to the same tweets. The Overlap metric captures the extent of such redundancy. Thus, given $t_i \in WC_k$ and $t_j \in WC_k$, we define the $Overlap(WC_k)$ of WC_k as:

$$\text{Overlap}(WC_k) = avg_{t_i \neq t_j} \frac{|Tw_{t_i} \cap Tw_{t_j}|}{\min\{|Tw_{t_i}|, |Tw_{t_j}|\}}, \tag{4}$$

If $Overlap(WC_k)$ is close to 0, then the intersections of tweets annotated by depicted words are small and such word clouds are more diverse.

Relevance of WC_k: Expresses how relevant the words in WC_k are to the query phrase t_q. We compute a relevance of a word cloud WC_k in the following fashion:

$$Relevance(WC_k) = avg_{t \in WC_k} \frac{|Tw_t \cap TwREL_{t_q}|}{|Tw_t|}, \tag{5}$$

The more Tw_t and $TwREL_{t_q}$ overlap, the more related t is to t_q. When $Tw_t \subseteq TwREL_{t_q}$, then t can be perceived as more specific sub-category of the original query t_q.

However, the Relevance measure does not capture ordering differences of words within the cloud and considers each term as a single query. The assumption that terms depicted in the cloud are of equal importance is often invalid. We believe that word weights and their order is an important aspect of word clouds where better ranked terms might be more visible i.e., larger font size or better position.

Further, we measure Mean Average Precision metric [5] for the evaluation of word clouds as follows:

We consider a generated word cloud as a query which should retrieve relevant tweets with respect to the query. Therefore, a better word cloud should link to more relevant tweets with respect to the query. We measure this as follows:

1. For given terms and corresponding weights of a word cloud WC_k, create a query vector Q_{WC_k} with normalized weights. Each entry of the query vector Q_{WC_k} represents the importance of a term from the word cloud WC_k with the normalized weight i.e., more important terms from the word cloud are represented with higher weights.
2. Rank and retrieve top-k tweets matching a given query Q_{WC_k}

3. Measure mean average precision(MAP) where each relevant tweet from TREC2011 microblog collection is considered a positive.

Ranking of relevant tweets with respect to a given query Q_{WC_k} is computed with standard information retrieval function OKAPI BM25 which can be defined as:

$$S(tw, Q_{WC_k}) = \sum_{q_i \in Q_{WC_k} \cap tw} c(q_i, Q_{WC_k}) \cdot TF(q_i, tw) \cdot IDF(q_i) \tag{6}$$

where

$$TF(q_i, tw) = \frac{f(q_i, tw) \cdot (k_1 + 1)}{f(q_i, tw) + k_1 \cdot (1 - b + b \cdot \frac{|tw|}{avgtwl})}$$

$$IDF(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}$$

and $f(q_i, tw)$ is a q_i term frequency within a tweet tw, $|tw|$ is the length of a given tweet tw, $avgtwl$ is average length of tweet within the corpus, N is a total number of tweets in the corpus and $n(q_i)$ is the number of tweets that contain the term q_i. To capture the importance of a word from the generated word cloud, we multiply the whole relevance score for a given term with the word cloud weight $c(q_i, Q_{WC_k})$ for the given term q_i. The function $c(q_i, Q_{WC_k})$ returns a weight of the term q_i from the query vector Q_{WC_k} which corresponds to the term weight from the word cloud WC_k. We set the same values for parameters $k_1 = 1.2$ and $b = 0.75$ as in [32].

We measured the average precision at K for the retrieved top K list of ranked tweets with respect to the given word cloud. Further, we measured the MAP for all generated word clouds. The average precision of top K ranked tweets with respect to the word cloud is calculated as follows:

$$AP@K(Q_{WC_k}) = \frac{\sum_k^K (P(k) \cdot rel(k))}{\#relevanttweets}$$

where $P(k)$ is the precision at k-th position in the ranked top K list and $rel(k)$ is 1 if the tweet at rank k is relevant, otherwise $rel(k)$ is 0 and $\#relevanttweets$ is the number of relevant tweets within the top K list. MAP is defined as:

$$MAP@K = \frac{\sum_{Q_{WC_k} \in AWC_k} AP@K(Q_{WC_k})}{|AWC_k|}$$

where AWC_k is the set of all generated word clouds and $AP@KQ_{WC_k}$ is average precision for the given word cloud Q_{WC_k}. In this work, we measure MAP at 30 under the assumption that it represents a reasonable cutoff for the number of relevant tweets similar to the approach in [30].

5.2 Baseline method

PageRank exploiting only extracted terms (PgRankTerms) This method was originally proposed in [3] to estimate tag relevance wrt. a certain query, and it outperformed

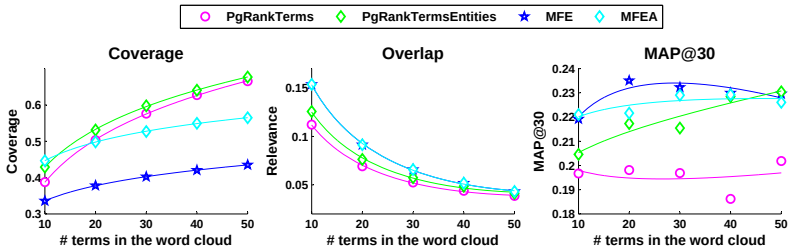

Fig. 2: Coverage, Overlap, and Mean Average Precision for word clouds of various sizes generated for queries from TREC2011 microblog collection.

several tag selection approaches in terms of relevance. In this work, the method estimates global terms importance within the graph created from the pooled tweets for the individual query from TREC2011 microblog collection. The β parameter is set to 0.85 (recommended value for a Pagerank algorithm). Due to the short nature of tweets, threshold α for edge creation between individual terms is set to 0. Shorter texts lead to small numbers of co-occurring terms, which consequently leads to a sparse graph.

5.3 Entity based methods

Most frequent entities (MFE) This method selects only recognized entities as defined in Section 3.1. The method provides a list of entities sorted by frequency in descending order, selecting top-*k* most popular entities.
Most frequent entities with grouped aliases (MFEA) This method selects only recognized entities and associated Freebase aliases as defined in Section 3.1. The method provides a list of entities sorted by frequency in descending order.
PageRank exploiting extracted terms, entities and grouped aliases (PgRankTermsEntities) This method estimates the global importance of terms and recognized named entities within the graph created from the extracted terms, recognized named entities and grouped Freebase aliases from pooled tweets for the individual query from TREC2011 microblog collection. The parameters are set to the same values as in the baseline method.

5.4 Results

We performed the evaluation on queries from TREC2011 microblog collection [30]. The MFE method has the worst Coverage ranging from 35% for word clouds with 10 terms to 45% for word clouds with 50 terms. MFEA has better Coverage with approximately 10% absolute improvement over the MFE method. The baseline method PgRankTerms attains greater Coverage than MFE and MFEA methods. The reason for higher Coverage of PgRankTerms is that entity mentions do not occur enough in tweets to outperform other extracted words.

However, when extracted words are combined with grouped named entities like in PgRankTermsEntities, the improvements in Coverage are highest. The PgRank-

TermsEntities method outperforms all other word cloud generation methods. PgRank-TermsEntities improves Coverage with respect to PgRankTerms and MFEA because it groups entity synonyms e.g. USA, US and America and represent them with the canonical entity name United States of America. In addition, it selects the most important terms which are not referring to named entities e.g., *#service*, *#jobs* for the query *BBC World Service staff cuts*. The relative improvements in comparison to PgRankTerms are 11% for 10 terms, 6% for 20 terms, 4% for 30 terms and 2% for 40 and 50 terms word clouds. Coverage improvements decrease as word clouds increase in size because the number of relevant/prominent recognized named entities in the underlying graph is lower. These results support the hypothesis H1: that grouping named entities improves the Coverage of word clouds.

Word cloud generation methods which exploit named recognized entities improve MAP. PgRankTermsEntities, MFE and MFEA outperform PgRankTerms in terms of MAP. The relative improvements of PgRankTermsEntities in comparison to PgRank-Terms are 4% for 10 terms, 10% for 20 terms, 9% for 30 terms, 23% for 40 and 14% for 50 terms word clouds. Thus, word clouds with named recognized entities improve access to the relevant tweets of the corpus which validates the H1 hypothesis. The main reason for the attained improvements is that almost 89% of all relevant tweets from TREC2011 microblog collection contain at least one recognized entity.

Similarly, 31% of all relevant tweets contain at least one Freebase alias (with minimal length of 4 characters). Comparing all pooled tweets from the TREC2011 microblog collection 77% contain recognized named entities and 28% of tweets contain at least one Freebase alias. Further, linking entity synonyms increases both Coverage and also the prominence of the named entity in the word cloud. Thus, it is more likely that the named entity will be represented in the word cloud and, if relevant for the query, it will improve access to the relevant tweets.

Improved access to relevant tweets and enhanced Coverage of word clouds can be attained through a combined selection of terms and recognized named entities. Thus, for enhanced word cloud generation it is important to combine recognized and grouped named entities with relevant and prominent terms from the underlying dataset.

The methods exploiting recognized named entities do have higher Overlap than the PgRankTerms method. We consider this finding interesting and unanticipated. The increased redundancies in the generated word clouds are caused by imperfect NER tools. In particular, tweets with an ambiguous name entity such as *BBC News Service* link to several semantically similar entities such as *BBC*, *BBC News*, *BBC NEWS Service*, which might lead to higher Overlap scores. Further, detected Freebase aliases might often increase Overlap for the similar reason e.g., alias *us* for *United States* covers many irrelevant tweets. To minimize the impact of ambiguous aliases we restrict the alias detection to a minimum length of 4 characters and the alias may not be a stop word.

Lemmatisation also had a positive effect on word cloud generation. Lemmatising terms to group them improves Coverage 1.75% above the baseline, and 3% for the PgRankTermsEntities. Similarly, MAP improves with an increase of 11% for PgRank-TermsEntities and 7% for the baseline technique. The negative impact of lemmatisation on word cloud generation is higher Overlap (decreased diversity of word clouds), with

an increase of 3% using the baseline technique. As the result is overall positive, we included lemmatisation as a preprocessing step for all cloud generation methods.

5.5 Diversification

To overcome the problems introduced by higher redundancy in word clouds, we investigate how to maximize global relevance as well as diversity of selected terms. Instead of following greedy diversification approaches, we take a unified approach of ranking global relevance together with the diversification objective. We use the DivRank algorithm [33] which assumes that transition probabilities change over time following the "rich gets richer" principle. The transition probability of visiting a node (term) A from other nodes is reinforced by the number of times a node A has been already visited. This reinforcement aspect is defined as $N_T(v)$ and it captures the number of visits to a node v until time T. Let us assume that at time T a random walk is at node u, then at time $T+1$ the walk proceeds to a node v with a transition probability $p_T(v|u)$ which is proportional to: $p(v|u) \cdot N_T(v)$. Hence, the general form of the DivRank algorithm can be declared as follows where a transition probability from a node u to node v at time T is:

$$p_T(v|u) = (1-\beta)\left(\frac{p_0(v|u)\cdot N_T(v)}{\sum_{v\in V} p_0(v|u)\cdot N_T(v)}\right) + \beta p_P$$

where p_P is the vector of prior probabilities and $p_0(v|u)$ is an initial estimation of transition probability which is equivalent to definition in Pagerank algorithm (see section 4.2).

It is challenging to approximate $N_T(v)$. A simple approximation might use $p_T(v)$ to estimate $N_T(v)$. Authors of DivRank algorithm [33] denote such an approximation as pointwise DivRank and is defined as follows:

$$p_T(v|u) = (1-\beta)\left(\frac{p_0(v|u)\cdot p_T(v)}{\sum_{v\in V} p_0(v|u)\cdot p_T(v)}\right) + \beta p_R$$

In the following, we present an impact of DivRank algorithm on the word cloud generation with grouped named entities. Figure 3 shows that with diversified word cloud generation, Overlap decreases. The relative improvements of DivRankTermsEntities outperforms the PgRankTerms baseline are 14% for 10 terms, 14% for 20 terms, 12% for 30 terms, 11% for 40 and 12% for 50 terms word clouds. The DivRankTermsEntities method significantly decreases Overlap in comparison to the PgRankTerms baseline (Wilcoxon signed-rank test, $p = 0.000094$) The improvements are even more significant with respect to PgRankTermsEntities method with 24% for 10 terms, 22% for 20 terms, 20% for 30 terms, 19% for 40 and 18% for 50 terms word clouds. In contrast, diversified word cloud generation significantly improves Coverage of word cloud generation. The improvement is statistically significant with respect to the baseline method PgRankTerms (Wilcoxon signed-rank test, $p = 0.0363$). The mean of relative improvements DivRankTermsEntities with respect to PgRankTermsEntities (the best performing method when measuring Coverage) is 2.35%.

Diversified word cloud generation from grouped and recognized named entities combined with extracted words decreases significantly Overlap, improves significantly Coverage and improves access to relevant tweets. This validates hypothesis $H1$.

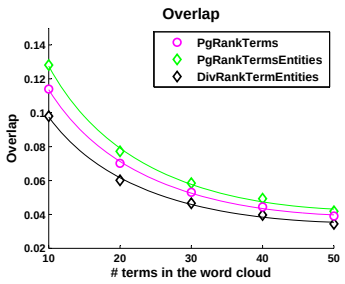

Fig. 3: Overlap for diversified word clouds with the method DivRankTermEntities of various sizes generated for queries from TREC2011 microblog collection.

5.6 Crowdsourced evaluation

In order to verify the findings from empirical evaluation of word clouds with different synthetic metrics, we designed a crowdsourced user evaluation of generated word clouds. We selected 8 queries from TREC2011 microblog collection for which we generated word clouds with DivRankTermEntities and PgRankTerms methods. We included 4 queries where the enhancement of MAP for word clouds with named entities with respect to the baseline was the greatest (denoted as Impr. MAP). Similarly, we added 4 word clouds for queries where the Overlap has been decreased the most with respect to the baseline (denoted as Impr. diversity ($\downarrow Overlap$)). The answers sought by the user evaluation are twofold. First, are word clouds with named entities perceived as more relevant and diverse by the end users? Second, do measured synthetic metrics correlate with the ratings of relevance and diversity by users?

Participants were asked to view a pair of word clouds, a set of tweets related to a certain query, and a related Wikipedia article. Their task was to determine which word cloud was more relevant and which was more diverse. The user was asked to rate the relevance and diversity of an individual word cloud with respect to the query on a Likert scale of 1 to 5 (Rating 1: word cloud A is very relevant/diverse to the pertaining query; Rating 3 - both word clouds are equally relevant/diverse to the pertaining query; and Rating 5 - word cloud B is very relevant/diverse to the pertaining query). We altered assignment of word clouds with named entities to either word cloud A or B for each query to prevent user bias that "word cloud A (with named entities) is always more relevant and diverse".

Non-grouped vs. Entity-grouped clouds Each word cloud pair was compared using 20 ratings from distinct users. For 7 out of 8 word clouds, the average ratings of relevance and diversity favoured word clouds generated with automatically grouped named

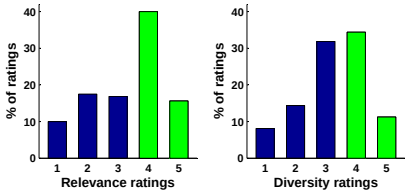

Fig. 4: Green bins (ratings 4 and 5) in the histograms indicate positive rating towards word clouds with grouped named entities. Ratings 1 and 2 indicate user preference towards the baseline word clouds and rating 3 represents that the baseline and the word cloud with grouped entities are equally relevant or diverse.

entities. For simplicity's sake, in the following we refer to word clouds generated with grouped entities as word cloud B; positive ratings are those over 3.0.

From 160 distinct relevance ratings, 89 were positive towards word clouds with named entities, 27 were neutral ratings and 44 were more towards the baseline generated word clouds (see Figure 4). Similarly for diversity ratings, 73 were positive towards word clouds with named entities, 51 were neutral ratings and 36 were more towards the baseline generated word clouds.

To further compare differences between word clouds generated by the baseline and clouds with grouped named entities, we performed a statistical significance test. The null hypothesis is that user ratings are normally distributed with mean 3.0, i.e., word clouds generated by the DivRankTermEntities and PgRankTerms methods are rated as equally relevant and equally diverse. For the relevance judgments, we found that word clouds generated by the DivRankTermEntities method are significantly better rated than the baseline word clouds ($p = 0.00062$, one sample t-test). Similarly, we determined that word clouds generated by the DivRankTermEntities are significantly better rated for diversity with respect to the baseline method ($p = 0.003$, one sample t-test). These findings support hypothesis $H2$: users find word clouds with grouped entities more relevant and diverse than those with no entity grouping.

Group	# clouds	min δ	mean δ
Impr. MAP	4	0.14	0.26
Impr. diversity ($\downarrow$ *Overlap*)	4	-0.02	-0.023
Decr. MAP & Overlap	2	-0.02	-0.133

Table 1: Three distinct groups statistics which were created according to the measured levels of synthetic metrics.

Synthetic metrics vs user perception The second goal of the user evaluation is to determine whether word clouds with higher levels of measured synthetic metrics are

rated by users as more relevant and diverse or vice versa. We focused on the MAP and Overlap metrics.[3]

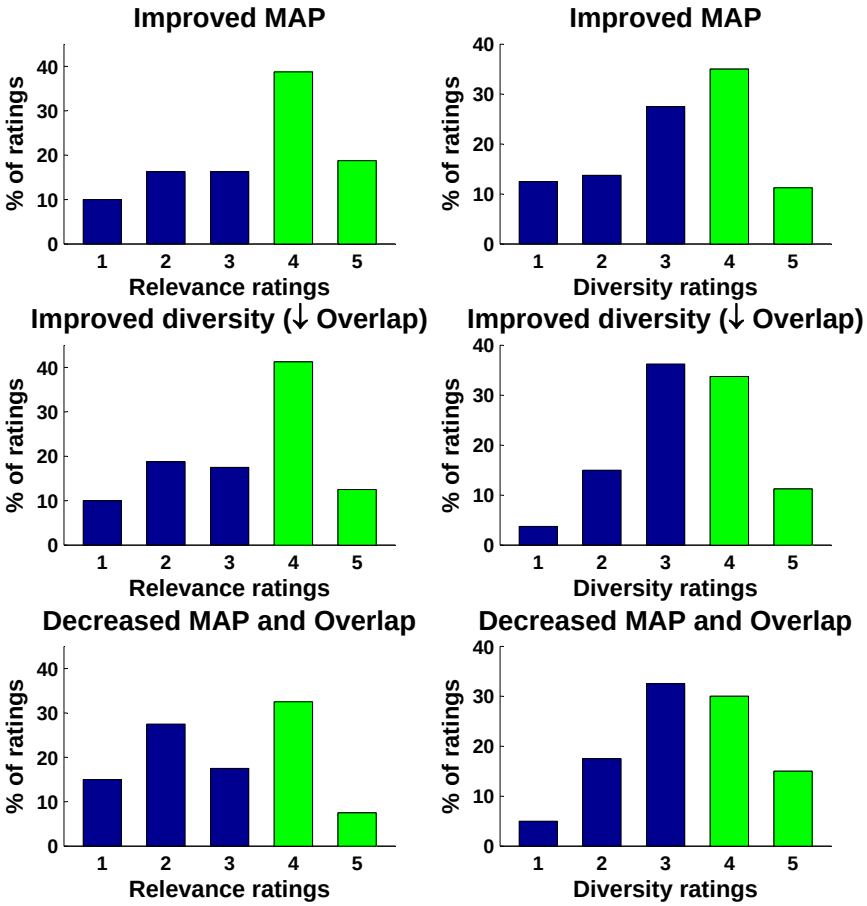

Fig. 5: Aggregated user ratings for three distinct groups of word clouds categorized according to the measured levels of synthetic metrics.

To determine the correlation between user judgments and synthetic metrics, we have created 3 different groups (see Table 1). We exploit the same two groups of word clouds Impr. MAP and Impr. diversity (↓ *Overlap*) as in Section 5.6. In addition, we added a group Decr. MAP & Overlap with two clouds where levels of MAP and Overlap were lower than the baseline word clouds. For each group, we report a minimum δ value which is a minimal difference between measured levels of the particular metric for

[3] Validation by users of the third metric introduced in [4], Coverage, is only possible with an interactive user evaluation. Hence, we do not include "coverage assessment" of word clouds in this study.

word clouds with grouped entities and the baseline. Hence a minimum δ is a threshold of measured synthetic metric whether to include a word cloud into the particular group. For instance, the threshold $\delta = 0.14$ for the Impr. MAP group indicates that only those word cloud pairs where the improvements of MAP are at least 0.14 (comparing the baseline and DivRankTermEntities methods) are included. The mean of δ expresses the average value of differences in metric values for each word cloud pair in the group, e.g., the average improvements of MAP in the group Impr. MAP is 0.26. Note that negative values of δ reflect cases where the metric is lower than baseline. For Decr. MAP & Overlap group, we only report levels of MAP due to substantial differences in comparison to the Overlap levels which have very slight differences between the baseline and the word clouds with grouped entities.

When all the ratings aggregated altogether from three groups, word clouds with grouped entities are still rated significantly more relevant ($p = 0.0046$, one sample t-test) and diverse ($p = 0.00047$, one sample t-test) than the baseline.

The relation between created groups and user judgments is presented in Figure 5. Users rated word clouds with higher MAP as more relevant. Of 80 ratings, 46 (57.5%) indicated that word clouds with grouped named entities are more relevant than the baseline. Conversely, for the word clouds with the decreased MAP and Overlap, only 40% of the ratings indicatie preference towards word clouds with grouped named entities. Hence, word clouds with higher MAP get 17.5% more positive ratings (4 or 5 ratings) than the baseline. The difference is even more pronounced for "rating 5 - much more relevant than the baseline word cloud", where Decr. MAP & Overlap group attained only 7.5% from all ratings, whereas the Improved MAP group attained 18.75%. Therefore, we can conclude that word clouds with grouped named entities which attain higher levels of MAP are more likely to be better rated in terms of relevance by users.

When measuring diversity, word clouds from the Impr. diversity($\downarrow$ *Overlap*) group were slightly more rated as "equally or more diverse than word clouds generated by the baseline" than other groups. In particular, with Impr. diversity($\downarrow$ *Overlap*), we observed a decreased number of ratings, expressing that the baseline word cloud is much more diverse (3.75% for Impr. diversity ($\downarrow$ *Overlap*) group and 12.5 for Impr. MAP). However, when looking at the decreased Map and Overlap group, the distribution of the ratings is fairly even. Hence, the Overlap metric is not a suitable predictor of user diversity ratings. This might be because the relative improvements of Overlap are too subtle to produce observable differences in user judgements of diversity. In order to attain more significant differences of Overlap, we believe that larger collections of tweets (retrieved w.r.t more general information needs) should be employed.

On the other hand, 46.3% of word clouds with improved MAP and 45% of word clouds from Decr. MAP & Overlap group were rated as more diverse than the baseline. Therefore, users rating word clouds with grouped entities have tend to find them more diverse than word clouds with no grouping.

6 Discussions, limitations and future work

False positives during entity recognition may have reduced relevant ratings. For instance, a word cloud generated for the query "*Super Bowl, seats*" contained "*Super*

(2010 American film)" which is irrelevant for this query. Similarly, for "*Kubica crash*", the entity "*crash bandicoot*" ended up in the word cloud. Unsupervised semantic clusterings (Section 3.2) may serve to better differentiate such cases.

Some word clouds generated with the PgRankTermsEntities suffered from increased Overlap. This was partially caused by imprecise named entity disambiguation where ambiguous named entities were not grounded correctly. Therefore, the quality of word clouds with grouped named entities is bounded by the precision of named entity annotation tools.

Evaluating word clouds with crowdsourced user evaluation is a challenging task due to uncertainty of reliability and quality of user ratings. In our pilot study, we aimed to ensure the quality of user ratings with pre-filtering quiz questions. However, we have observed that for test questions where users were asked to rate word cloud diversity (one cloud was supposed to be more diverse) many participants disagreed. Due to the subjective nature of the task, we disregarded a user "qualifying" phase (as is often best practice in crowdsourcing [34]) and instead aimed to collect more user ratings and observe aggregated ratings. To further ensure the quality of the ratings, we accepted ratings only from participants in English-speaking countries, as word clouds were generated from tweets written in English.

As future work, we would like to design a hybrid clustering method which would combine knowledge from linked data repositories (e.g., Freebase) with probabilistic context of terms (the similar intuition as in Brown clustering). The hybrid approach could further improve accuracy of entity grouping as context of terms could minimize incorrect grouping of aliases. Further, the future work would explore how grouped named entities could improve a personalized word cloud generation [5], mainly whether grouped named entities could alleviate a sparsity problem when expressing user preferences.

7 Conclusion

Generating word clouds from social streams is a difficult task. Because users often discuss the same entity using multiple aliases, the utility of word clouds becomes degrades when this complex and high-variety data is used directly. Consequently, methods of unifying these variations become necessary, in order to get accurate counts. Accordingly, we propose techniques for grouping aliases that refer the same entity, and for representing these groups using a canonical term. The method improves the coverage of word clouds and access to the relevant content.

This variety also leads to redundant terms, that must be clustered together, in order to improve the precision of the cloud. Due to imperfect state-of-the-art named entity recognition on social media, redundancy of terms in word clouds often remains. This makes it necessary diversify terms. We found that unsupervised term extraction and clustering techniques (such as Brown clustering) can be used to automatically identify similar and co-referent terms, beyond the lists available through commercial and third-party ontology services. It was then demonstrated that our technique not only significantly decreases redundancy, but also leads to significantly higher coverage than baseline word cloud generation, leading to better word clouds and therefore improved

information access. Combined, these factors alleviate problems with in clouds from social media.

Naturally, this leads to questions about how evaluation can be tested. Earlier, we hypothesised that word clouds with grouped named entities are significantly more relevant and diverse than word clouds with no entity grouping. This was evaluated extrinsically against the crowd, with reported user experiences supporting the hypothesis. Further, word clouds with grouped named entities that score higher MAP are more likely to be rated as relevant by users.

Finally, we compared these gold-standard human judgments to a proposed synthetic cloud evaluation metric. It was shown that this previously-proposed MAP metric for automatic cloud evaluation predicts extrinsic human evaluations of cloud quality. Thus, when designing word clouds, the MAP metric should be used as a quality predictor of the cloud generation technique, enabling automatic assessment of word cloud quality without a human in the loop.

Acknowledgments This work was partially supported by the European Union under grant agreement No. 611233 PHEME.

References

1. Kuo, B.Y., Hentrich, T., Good, B.M., Wilkinson, M.D.: Tag clouds for summarizing web search results. In: Proceedings of the conference on the World Wide Web (WWW), ACM (2007) 1203–1204
2. Miotto, R., Jiang, S., Weng, C.: eTACTS: A method for dynamically filtering clinical trial search results. Journal of Biomedical Informatics **46**(6) (2013) 1060–1067
3. Leginus, M., Dolog, P., Lage, R.: Graph based techniques for tag cloud generation. In: Proceedings of the ACM Conference on Hypertext and Social Media, ACM (2013) 148–157
4. Venetis, P., Koutrika, G., Garcia-Molina, H.: On the selection of tags for tag clouds. In: Proceedings of the conference on Web Search and Data Mining (WSDM), ACM (2011) 835–844
5. Leginus, M., Zhai, C., Dolog, P.: Personalized generation of word clouds from tweets. Journal of the Association for Information Science and Technology (2015)
6. Tufekci, Z.: Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In: Proceedings of the International Conference on Weblogs and Social Media (ICWSM), AAAI (2014) 505–514
7. Bernstein, M.S., Suh, B., Hong, L., Chen, J., Kairam, S., Chi, E.H.: Eddi: interactive topic-based browsing of social status streams. In: Proceedings of the annual symposium on User Interface Software and Technology (UIST), ACM (2010) 303–312
8. Lage, R., Dolog, P., Leginus, M.: The role of adaptive elements in web-based surveillance system user interfaces. In Dimitrova, V., Kuflik, T., Chin, D., Ricci, F., Dolog, P., Houben, G.J., eds.: User Modeling, Adaptation, and Personalization (UMAP). Volume 8538 of Lecture Notes in Computer Science. Springer International Publishing (2014) 350–362
9. Rout, D., Bontcheva, K., Hepple, M.: Reliably evaluating summaries of Twitter timelines. In: Proceedings of the AAAI Workshop on Analyzing Microtext, AAAI (2013) 64–71
10. Derczynski, L., Maynard, D., Aswani, N., Bontcheva, K.: Microblog-genre noise and impact on semantic annotation accuracy. In: Proceedings of the ACM Conference on Hypertext and Social Media, ACM (2013) 21–30

11. Maynard, D., Greenwood, M.A.: Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. In: Proceedings of the conference on Language Resources and Evaluation (LREC), Reykjavik, Iceland, ELRA (2014)
12. Leginus, M., Derczynski, L., Dolog, P.: Enhanced information access to social streams through word clouds with entity grouping. In: Proceedings of the conference on Web Information Systems and Technologies (WEBIST). (2015)
13. Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., Dredze, M.: Annotating named entities in twitter data with crowdsourcing. In: Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, ACL (2010) 80–88
14. Hogan, A., Zimmermann, A., Umbrich, J., Polleres, A., Decker, S.: Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora. Web Semantics: Science, Services and Agents on the World Wide Web **10** (2012) 76–110
15. Hu, Y., Talamadupula, K., Kambhampati, S., et al.: Dude, srsly?: The surprisingly formal nature of Twitter's language. In: Proceedings of the International Conference on Weblogs and Social Media (ICWSM), AAAI (2013)
16. Baldwin, T., Cook, P., Lui, M., MacKinlay, A., Wang, L.: How noisy social media text, how diffrnt social media sources. In: Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP). (2013) 356–364
17. Bontcheva, K., Derczynski, L., Funk, A., Greenwood, M.A., Maynard, D., Aswani, N.: TwitIE: An open-source information extraction pipeline for microblog text. In: Proceedings of the conference on Recent Advances in Natural Language Processing (RANLP). (2013) 83–90
18. Baldwin, T., Kim, Y.B., de Marneffe, M.C., Ritter, A., Han, B., Xu, W.: Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. ACL-IJCNLP 2015 (2015) 126
19. Han, B., Baldwin, T.: Lexical normalisation of short text messages: Makn sens a #twitter. In: Proceedings of the annual meeting of the Association for Computational Linguistics (ACL), ACL (2011) 368–378
20. Augenstein, I., Gentile, A.L., Norton, B., Zhang, Z., Ciravegna, F.: Mapping keywords to linked data resources for automatic query expansion. In: Proceedings of the Second International Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data. (2013) 9–20
21. Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., Petrak, J., Bontcheva, K.: Analysis of named entity recognition and linking for tweets. Information Processing & Management **51**(2) (2015) 32–49
22. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the meeting of the Special Interest Group on Management of Data (SIGMOD), ACM (2008) 1247–1250
23. Kergl, D., Roedler, R., Seeber, S.: On the endogenesis of Twitter's Spritzer and Gardenhose sample streams. In: Proceedings of the conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE (2014) 357–364
24. Brown, P.F., Desouza, P.V., Mercer, R.L., Pietra, V.J.D., Lai, J.C.: Class-based n-gram models of natural language. Computational linguistics **18**(4) (1992) 467–479
25. Wittgenstein, L.: Philosophical investigations. Basic Blackwell, London (1953)
26. Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: Proceedings of the annual international conference on Systems documentation (SIGDOC), ACM (1986) 24–26
27. Derczynski, L., Chester, S., Bøgh, K.S.: Tune Your Brown Clustering, Please. In: Proceedings of the conference on Recent Advances in Natural Lang Processing (RANLP). (2015)

28. Lui, M., Baldwin, T.: langid.py: An off-the-shelf language identification tool. In: Proceedings of the annual meeting of the Association for Computational Linguistics (ACL). Volume 3., ACL (2012) 25–30
29. Wu, W., Zhang, B., Ostendorf, M.: Automatic generation of personalized annotation tags for Twitter users. In: Proceedings of the annual meeting of the Association for Computational Linguistics (ACL), ACL (2010) 689–692
30. Ounis, I., Macdonald, C., Lin, J., Soboroff, I.: Overview of the TREC-2011 microblog track. In: Proceedings of the Text REtrieval Conference (TREC). (2011)
31. McCreadie, R., Soboroff, I., Lin, J., Macdonald, C., Ounis, I., McCullough, D.: On building a reusable Twitter corpus. In: Proceedings of the meeting of the Special Interest Group in Information Retrieval (SIGIR), ACM (2012) 1113–1114
32. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Volume 1. Cambridge University Press (2008)
33. Mei, Q., Guo, J., Radev, D.: Divrank: the interplay of prestige and diversity in information networks. In: Proceedings of the meeting of the Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD), ACM (2010) 1009–1018
34. Sabou, M., Bontcheva, K., Derczynski, L., Scharl, A.: Corpus annotation through crowdsourcing: Towards best practice guidelines. In: Proceedings of the conference on Language Resources and Evaluation (LREC), ELRA (2014)