Semantic Linking and Contextualization for Social Forensic Text Analysis

Zhaochun Ren, David van Dijk, David Graus, Nina van der Knaap, Hans Henseler, Maarten de Rijke

ISLA, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands
z.ren@uva.nl, d.v.van.dijk@hva.nl, d.p.graus@uva.nl,
n.van.der.knaap@law.leidenuniv.nl, j.henseler@hva.nl, derijke@uva.nl

***Abstract*—With the development of social media, forensic text analysis is becoming more and more challenging as forensic analysts have begun to include this information source in their practice. In this paper, we report on our recent work related to semantic search in e-discovery and propose the use of entity and topic extraction for social media text analysis. We first describe our approach for entity linking at the 2012 Text Analysis Conference Knowledge Base Population track and then detail the personalized tweets summarization task is introduced, where entity linking is used for semantically enriching information in a social media context.**

Keywords*—Entity linking; Tweets summarization; Forensic text analysis; Semantic search in e-discovery;

I. Introduction

Forensic text analysis focusses on extraction and analysis of content from data for crime investigation. Social media has become an invaluable source of information for this task. In social media applications, users broadcast and propagate news, stories and burst events, or buzz their own sentiment comments that reflect their biased attitude to some event or topic. Analysts can, e.g., try to infer user information, such as personal characteristics, interests or educational background through content analysis.

But extracting meaningful information from social media documents or social network data collections turns out to be quite a challenge. The data differs to "classical" document sets in many ways which gives new challenges. For example, most traditional concept extraction methods assume source documents to be relatively clean and grammatically correct, but many social media documents are short and ungrammatical. Also the information overload problem and missing information are a challenge for social media analysts, because of the time-aware nature and large data volume. How to recognize whether something is important in these data streams is of great research interest.

We believe semantic search will help to answer this question. Semantic search is a paradigm in Information Retrieval (IR) which applies structured knowledge, e.g., discussion structure, topical structure, or entities and relations, as a complement to text retrieval. In this context we sketch our recent and ongoing work. We focus on the task of semantic linking and contextualization for social forensic text analysis: (i) Firstly, we introduce our previous work on the 2012 Text Analysis Conference Knowledge Base Population task [1] that handles the task of entity linking. Entity linking addresses the problem of disambiguating entity mentions in unstructured text against a background knowledge base. In our method, we adapt a state-of-the-art entity linking method [2] for micro-blog posts, which links entity mentions in micro-blog posts to relevant Wikipedia articles. (ii) Based on our entity linking results, we have studied personalized tweets summarization [3]. To remedy the length limitation problem of each tweet, we select most salient sentences from the linked Wikipedia article and put them in the tweet. Then, a personalized time-aware strategy for selecting tweets is proposed using a single, unified topic model [4].

The rest of this paper is organized as follows. Our proposed strategies for semantic linking and contextualization are detailed in Section II. Section III presents and discusses our experimental results and Section IV concludes the paper.

II. Entity Linking and Tweets Summarization

In this section, we describe our previous work on entity linking (II-A) and tweets summarization (II-B) respectively. We use entity linking to identify entities in social media and then link them with Wikipedia articles. We employ personalized time-aware tweets summarization based on a users history and collaborative social influences from "social circles."

A. Content-Based Entity Linking

The entity linking task can be formally described as follows: given an entity mention q_r (a term or phrase) occurring in reference document r, identify the entity e from a knowledge base KB that is the most likely referent of q_r. In our entity linking approach, we view linking as a binary classification task: given a mention q_r and entity candidate c, determine whether q_r refers to c or not. Our approach consists of the following steps: (i) candidate generation, (ii) candidate disambiguation (or re-ranking) and finally (iii) NIL detection and clustering, where we assign unique ids to queries that refer to entities that are not represented in the Knowledge Base.

In the candidate generation phase, given a query, we use the entity mention as input for a search over Wikipedia article titles, disambiguation pages and anchors to return the disambiguation candidates. For the disambiguation step, we apply a machine learning approach, and provide three feature sets and combinations thereof:

1) **Baseline features**: we apply a subset of the features in [2], which involve strictly the query and its candidate. The features include measures such as similarity between

query and candidate title, edit distance between the two, and structural properties, such as the number of links to or from the candidate entity. In total, 32 baseline features are used in our work.

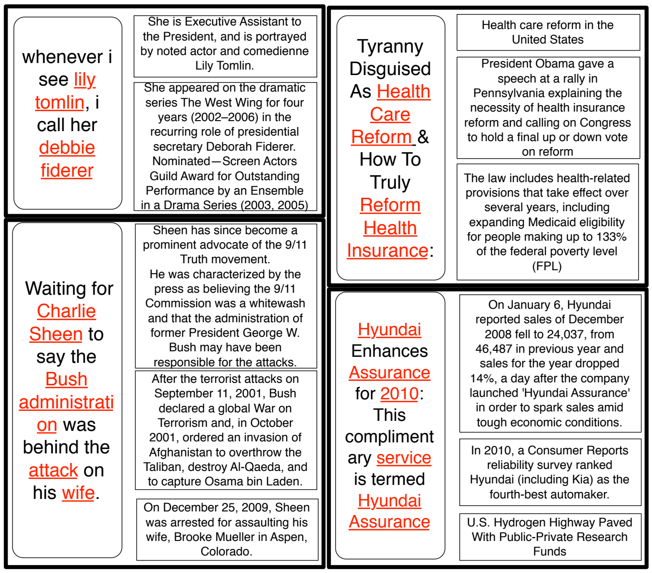
2) **Context features**: The context features make use of the semantic information encapsulated in the graph structure of the knowledge base: in this graph structure, we consider entities nodes, and links between them edges. Context features are based on the presence of related entities to the candidate entity. We perform a search for related entities in the document r, and derive features from several properties of the document, e.g. the number of titles and anchors of related entities found in the document, and common statistics associated with the retrieved anchors. In total, 40 context features are employed.

3) **LOD features**: We adopt the approach of [5], and perform joint disambiguation by considering all possible entity candidates in the set of entity mentions from document r simultaneously. We use the Linked Open Data cloud to obtain vector representations of entities and the document, as it provides a richer source of structured data than Wikipedia. We find the most similar entity candidate by measuring similarity between the document vector and the candidate vector.

B. Personalized Tweets Summarization

As the second subtask, we focus on the task of personalized time-aware tweets selection. Our approach to the task is based on the entity linking task that connects entities to Wikipedia articles. Since short and ungrammatical tweets hinder the forensic analysis, our remedial strategy is to expand the original tweet by adding relevant information from linked Wikipedia articles. After obtaining the three most likely Wikipedia articles, we extract the most central sentences from these Wikipedia articles and append them to the tweet. Figure 1 shows an example of this type of document expansion.

Most Twitter users rarely post tweets of their own, thus intuitions from the field of collaborative filtering is considered. Given two users u_i and u_j on Twitter, there are two main reasons for u_i and u_j to follow each other: either because they have similar interests or they have some relationship outside Twitter [6]. We define a *social circle* around a user u to be a set of friends of u such that every pair of users in this set follow each other on Twitter. See Figure 2 for a schematic representation. We assume that each Twitter user's interests are represented by a multinomial distribution $\theta_{u,t}$, which may, however, change over time. That is, the time-aware interests of user u are represented as a multinomial distribution $\theta_{u,t}$ over topics, where each topic is represented as a probabilistic distribution over words [4]. Formally, we have $\theta_{u,t} = \{\theta_{u,t,z_1}, \cdots, \theta_{u,t,z_K}\}$, where θ_{u,t,z_i}, denotes the distribution of topic z_i for user u at time t. We assume that topic distributions are dynamic and may differ between time periods. Posterior distributions over those parameters are derived by a Gibbs EM sampling algorithm [7].



whenever i see lily tomlin, i call her debbie fiderer

She is Executive Assistant to the President, and is portrayed by noted actor and comedienne Lily Tomlin.

She appeared on the dramatic series The West Wing for four years (2002–2006) in the recurring role of presidential secretary Deborah Fiderer. Nominated—Screen Actors Guild Award for Outstanding Performance by an Ensemble in a Drama Series (2003, 2005)

Waiting for Charlie Sheen to say the Bush administration was behind the attack on his wife.

Sheen has since become a prominent advocate of the 9/11 Truth movement. He was characterized by the press as believing the 9/11 Commission was a whitewash and that the administration of former President George W. Bush may have been responsible for the attacks.

After the terrorist attacks on September 11, 2001, Bush declared a global War on Terrorism and, in October 2001, ordered an invasion of Afghanistan to overthrow the Taliban, destroy Al-Qaeda, and to capture Osama bin Laden.

On December 25, 2009, Sheen was arrested for assaulting his wife, Brooke Mueller in Aspen, Colorado.

Tyranny Disguised As Health Care Reform & How To Truly Reform Health Insurance:

Health care reform in the United States

President Obama gave a speech at a rally in Pennsylvania explaining the necessity of health insurance reform and calling on Congress to hold a final up or down vote on reform

The law includes health-related provisions that take effect over several years, including expanding Medicaid eligibility for people making up to 133% of the federal poverty level (FPL)

Hyundai Enhances Assurance for 2010: This compliment ary service is termed Hyundai Assurance

On January 6, Hyundai reported sales of December 2008 fell to 24,037, from 46,487 in previous year and sales for the year dropped 14%, a day after the company launched 'Hyundai Assurance' in order to spark sales amid tough economic conditions.

In 2010, a Consumer Reports reliability survey ranked Hyundai (including Kia) as the fourth-best automaker.

U.S. Hydrogen Highway Paved With Public-Private Research Funds

Fig. 1. Four examples for Entity Linking and Ranking corresponding to four individual tweets, where the text box on the left-hand side indicates the original tweet and the text box on the right-hand side shows the extracted related sentences.

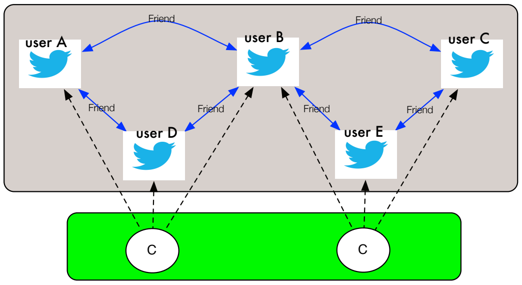


Fig. 2. An example of social circles on Twitter: there are two social circles (indicated using the 'c') among the five users in this graph, where each pair of vertices in each social circle is connected through the "friend" relationship.

Typically, traditional summarization does not cover the evolution of a specific event. Given a split of a user's history into time periods, the task of time-aware tweets summarization is to select the most representative tweets for each time period, covering the whole event evolution on a timeline. More precisely, given a set of tweets $\mathcal{D}$, a set of time periods T, and a maximum number of tweets per period, N, time-aware tweets summarization aims to extract multiple sets of tweets RT_t $(1 \leq t \leq T)$ from $\mathcal{D}$, where for each time period t, RT_t is a set of representative tweets $RT_t = \{d_{t,x_1}, d_{t,x_2}, \ldots, d_{t,x_N}\}$ that summarize the period. Furthermore, *personalized* time-aware tweets summarization is defined similar to time-aware tweets summarization, but in this case the tweets selected for inclusion in RT_t need to be relevant based on u's interests θ_u at time t.

III. Experiment

We evaluate semantic linking step by means of the setup provided by the Text Analysis Conference Knowledge Base Population evaluation campaign. We present our results in Table I. and report the B-cubed+ F1 scores for the full set of queries in the 2012 campaign, the subset of in-KB queries (non NIL) and the subset of NIL queries.

As we show in Table I, our NIL clustering improves overall performance when fewer entities are linked. As our NIL approach was identical across the five runs, the distinction between runs in the in-KB subset provide more valuable insights. In this subset, combining the baseline with the context extension achieves highest performance. We believe this is caused by the datasets' ambiguity: single query mentions can refer to multiple entities, and multiple queries can refer to a single entity. Our baseline approach links identical mentions to the same entity, so it does not cope well with this ambiguity.

TABLE I. Bˆ3+ F1 Results for Linking.

Full query set (2226)	official	corrected
Baseline (BL)	0.379	0.387
Context (Co)	0.428	0.427
BL+Co	**0.450**	0.434
BL+LOD	0.399	0.383
BL+Co+LOD	0.437	0.428
NIL subset (1049)	official	corrected
Baseline (BL)	0.388	0.398
Context (Co)	**0.648**	0.493
BL+Co	0.493	0.445
BL+LOD	0.446	0.399
BL+Co+LOD	0.469	0.434
In-KB subset (1177)	official	corrected
Baseline (BL)	0.364	0.370
Context (Co)	0.231	0.364
BL+Co	0.407	**0.418**
BL+LOD	0.351	0.361
BL+Co+LOD	0.402	0.415

For the task of personalized tweets summarization, a Twitter dataset that includes both social relations and tweets is used. It contains 47,373,408 tweets published by 562,361 users in 2009, and 295,145,421 tweets published by 3,153,356 users in 2010. Table II shows the average performance of our strategies and baselines, in terms of ROUGE-1, ROUGE-2 and ROUGE-W, based on all candidate tweets in all time periods. We find that our method outperforms the baselines in every case. "TPM-based runs" in Table II refers to our proposed methods: TPM-ALL refers to the combined model; TPM-SOC to the model that only considers users social influence and TPM-TOP to the model that uses a users social circles.

We evaluated our performance for a varying number N of tweets selected per period. As shown in Table II, TPM-ALL performs much better than other baselines. For $N = 40$, TPM-ALL achieves an increase of 10.6%, 11.6% and 8.9% over MM-AT in terms of ROUGE-1, ROUGE-2, and ROUGE-W respectively. For the dynamic version without social influence, TPM-TOP and TPM-SOC outperforms all other baselines also, which indicates the effectiveness of detecting dynamic topics.

TABLE II. Overall Summarization Performance

Metrics	TPM-ALL	TPM-TOP	TPM-SOC	MM-AT	T-LDA
Cut-off of $N = 40$ tweets per period					
ROUGE-1	**0.428**	0.403	0.395	0.387	0.374
ROUGE-2	**0.125**	0.119	0.116	0.112	0.112
ROUGE-W	**0.159**	0.153	0.149	0.146	0.142
Cut-off of $N = 60$ tweets per period					
ROUGE-1	**0.513**	0.497	0.482	0.461	0.457
ROUGE-2	**0.149**	0.143	0.139	0.134	0.127
ROUGE-W	**0.197**	0.191	0.189	0.178	0.176

IV. Outlook

In this paper, we described our previous work about entity linking and tweets summarization. Our previous work uses Wikipedia as a knowledge base for entity linking; however, criminal person entities are typically not listed there. Using the approach in [8], a knowledge base of identities relevant to the case under investigation can be compiled from seized evidence. We aim to plug this approach into our future content-based entity linking work.

Our work on entity linking and personalized tweets summarization forms the start of our semantic search in e-discovery project [9]. At its heart, e-discovery is the practice of sense making in textual corpora. By combining expertise from the fields of law and criminology with that of information retrieval and extraction, we aim to move beyond "algorithm-centric" evaluation, towards evaluating the impact of semantic search in real search settings. We to approach this through collaboration in an interdisciplinary group, consisting of four Phd candidates, one of whom is focusing on users and use case analysis, one on system development and two on algorithm development.

In our methodology, we apply an iterative two-phase work cycle within four sub-projects that run in parallel. During the first phase we work individually. We determine the use of, and needs for, intelligent search technology in e-discovery, and simultaneously explore and develop state-of-the-art semantic search approaches, as sketched in this note. In the second phase we collaborate, designing user experiments to evaluate how and where semantic search can support the analysts search process.

Acknowledgements

This research was partially supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements nr 258191 (PROMISE Network of Excellence) and 288024 (LiMoSINe project), the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.061.814, 612.061.815, 640.004.802, 727.011.005, 612.001.116, HOR-11-10, the Center for Creation, Content and Technology (CCCT), the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, the WAHSP and BILAND projects funded by the CLARIN-nl program, the Dutch national program COMMIT, by the ESF Research Network Program ELIAS, Elite Network Shifts project funded by the Royal Dutch Academy of Sciences, the eLaw group of Leiden University's Institute for the Interdisciplinary Study of the Law, and Fox-IT.

References

[1] D. Graus, T. Kenter, M. Bron, E. Meij, and M. de Rijke, "Context-based entity linking – The University of Amsterdam at TAC 2012," in *TAC 2012*, 2012.

[2] E. Meij, W. Weerkamp, and M. de Rijke, "Adding semantics to microblog posts," in *WSDM 2012*, 2012.

[3] Z. Ren, S. Liang, E. Meij, and M. de Rijke, "Personalized time-aware tweets summarization," in *SIGIR 2013*, 2013.

[4] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning research*, vol. 3, pp. 993–1022, 2003.

[5] S. Cucerzan, “Large-scale named entity disambiguation based on Wikipedia data,” in *EMNLP ’07*. ACL, 2007, pp. 708–716.

[6] Z. Xu, Y. Zhang, Y. Wu, and Q. Yang, “Modeling user posting behavior on social media,” in *SIGIR 2012*, 2012, pp. 545–554.

[7] H. Wallach, “Topic modeling: beyond bag-of-words,” in *ICML 2006*, 2006, pp. 977–984.

[8] J. Hofste, H. Henseler, and M. van Keulen, “Computer assisted extraction, merging and correlation of identities with tracks inspector,” in *ICAIL 2013*, 2013.

[9] D. van Dijk, H. Henseler, and M. de Rijke, “Semantic search in e-discovery,” in *DESI IV Workshop on Setting Standards for Searching Electronically Stored Information In Discovery Proceedings*, 2011.