Information extraction from social media: A linguistically motivated approach

Nelleke Oostdijk[1], Ali Hürriyetoglu[1], Marco Puts[2],
Piet Daas[2], Antal van den Bosch[1]
(1) Centre for Language Studies, Radboud University, P.O. Box 9103, NL-6500 HD, Nijmegen, the Netherlands
(2) Statistics Netherlands, P.O. Box 4481 6401 CZ, Heerlen, the Netherlands
n.Oostdijk@let.ru.nl, a.hurriyetoglu@let.ru.nl, m.puts@cbs.nl,
pjh.daas@cbs.nl, a.vandenbosch@let.ru.nl

RÉSUMÉ

Extraction d'information des réseaux sociaux : une approche motivée linguistiquement
Nous proposons une méthode flexible pour l'extraction de l'information sur le trafic à partir des réseaux sociaux. L'abondance de micropostes sur le Twitter rend possible d'identifier ce qui se passe parce que les utilisateurs rapportent ce qu'ils sont en train d'observer. Cette information est très pertinente et peut aider les organisateurs de la sécurité routière et les conducteurs à être mieux préparés et d'effectuer les actions appropriées. Nous distinguons 22 catégories d'information supposées être pertinentes au domaine du trafic. Nous atteignons alors un score de 74% avec les tweets individuels. Nous jugeons cette performance satisfaisante, d'autant plus qu'il existe habituellement plusieurs tweets sur un événement donné, ce qui nous permet de détecter l'information pertinente.

ABSTRACT

Information extraction from the social media : a linguistically motivated approach

We propose a flexible method for extracting traffic information from social media. The abundance of microposts on Twitter make it possible to tap into what is going on as users are reporting on what they are actually observing. This information is highly relevant as it can help traffic security organizations and drivers to be better prepared and take appropriate action. Distinguishing 22 information categories deemed relevant to the traffic domain, we achieve a success rate of 74% when individual tweets are considered. This performance we judge to be satisfactory, seeing that there are usually multiple tweets about a given event so that we will pick up what relevant information is out there.

MOTS-CLÉS : fouille de réseaux sociaux, extraction d'information, information sur le trafic, sécurité routière.

KEYWORDS: social media mining, information extraction, traffic information, traffic safety.

1 Introduction

Accurate real-time traffic information is highly relevant for a number of reasons, ranging from economic interests that are at stake to safety issues. Therefore, governments invest substantially in measuring and forecasting traffic. They set up expensive on-road sensors, or track vehicles to enable traffic information to be collected as efficiently and quickly as possible (Leduc, 2008). However, extracting actionable insights from this data can be time-consuming while requiring expert knowledge,

and data may be too sparse to reliably extract patterns from. Therefore, we introduce social media, specifically Twitter, as an additional rich source of relatively explicit and real-time traffic information to support drivers and decision makers.

On Twitter, users will report on what they observe as they are driving along and give information as regards the flow of traffic, accidents occurring, adverse weather conditions, etc. In addition, Twitter users share their expectations about or speculate on how the current situation will develop. A detailed analysis of such tweets may provide many details that are not specified in or have not yet been identified by traditional traffic information systems.

Mining tweets for any time period and for any region is relatively feasible. Traffic-related tweets may be filtered from the Twitter Streaming API [1] in real time to monitor the traffic flow, or may be searched offline to gather additional information about what actually transpired before and after a recorded incident.

With this objective in mind, we developed a rule-based formal information extraction methodology that suits the flexibility of the language used on social media. We specified a Backus Naur Form (BNF) grammar which uses a hierarchical set of rules and a base lexicon containing known lexical items that are relevant to the traffic domain. The grammar was implemented using the Pyparsing Library (McGuire, 2007).

We applied our method to social media data to identify traffic-related information in tweets, distinguishing between 22 information categories. We ultimately aim at automatically constructing a database of traffic information. This database will be used to report the traffic flow in real time, and query the collected information to identify relations among traffic events retrospectively. The complete system may improve traffic intelligence and safety, support the prediction of upcoming traffic events, and enhance compiling periodic traffic reports [2].

The extracted information (possibly in combination with data from other sources) can be explored and exploited so as to

— gain insight in the frequency of occurrence of various events (accidents, traffic jams, roads being closed, etc.) ;
— discover correlations ; for example, how often, over a given period of time, were there reports of traffic jams on a specific highway which also mentioned accidents or ongoing road works ;
— explore relations among event types ; for instance : a weather event may more likely cause a traffic event at some places than at other places ;
— identify information about relatively small events mentioned in only a few tweets ; one tweet from which information can be accurately extracted should already be enough to detect a traffic event.

We discuss related research in Section 2. We introduce the information extraction method in Section 3, the data used in our experiments in Section 4, the information categories about traffic in Section 5, and the results in Section 6. The source code of the implementation is available. [3]

1. https://dev.twitter.com/streaming/public

2. We work with Statistics Netherlands on improving periodic traffic report compilation. Moreover, we think the demo of the proposed system, which is on http ://sinfex.science.ru.nl, can already enhance drivers' understanding of the traffic in the Netherlands.

3. The source code of the implementation, tweet IDs of the training data, and the annotated test tweets used for this paper can be accessed on Bitbucket at https://bitbucket.org/hurrial/sinfex/commits/c50cc5b44cd0d8a4da3fc6303596f2ebefa2b745.

2 Related Work

The richness of the content found on social media has caused researchers to seek ways of utilizing this source for many specific domains, including the traffic domain (Gal-Tzur *et al.*, 2015). Huzita *et al.* (2012), Tostes *et al.* (2014), and Cui *et al.* (2014) used Facebook [4], Foursquare [5] and Instagram [6], and Sina Weibo [7] data respectively to identify traffic conditions.

Twitter data analysis has become a popular type of streaming social media mining due to the openness of this platform. Studies in this line are mainly about traffic-related tweet detection (D'Andrea *et al.*, 2015), traffic situation prediction (He *et al.*, 2013), and traffic forecasting (Wibisono *et al.*, 2012). The utility of the social aspect in traffic information sharing is also used by dedicated applications such as Waze [8] which provides a platform to share traffic events with other users of this platform.

The analysis methods vary from general machine learning (Axel Schulz *et al.*, 2013; Anantharam *et al.*, 2014; Wanichayapong Napong, Pattara-Atikom Wasan & Ratchata, 2014) to detailed language analysis techniques (Endarnoto *et al.*, 2011). Our approach falls in the latter category. We aim to handle the richness of social media language while our method should also be able to identify small events. The information extraction method that we use can be characterized as pattern-based, hierarchical, information rich, and adaptive. It is explained in Section 3.

3 Method

We developed a robust and modular information extraction method that can handle text on social media and in any language effectively, provided that rules for that language are written. The domain knowledge, place names, [9] temporal expressions, [10] and specific lexical patterns are provided as language resources. The method uses a training set to learn related lexical items such as place names via high-precision linguistic patterns. Using these resources, we wrote a modular BNF grammar which allows for some degree of flexibility and robustness by allowing tokens to match partially or allow the '#' character at the beginning of a token.

The definition of the BNF grammar starts with specifying acceptable tokens, for instance a token may contain a hash '#' at the beginning, a dot '.' at the end, or an apostrophe or a dash in the middle. We allow case-insensitive match for the sake of flexibility. Next, we specify whether domain terms should allow variation at the beginning or at the end, e.g., *route*, *tunnel*. For example, allowing flexible matching at the beginning of the term *file* (EN : 'traffic jam') enables this token to represent any token that ends with this term, e.g., *kijkersfile*, *windmolenkijkersfile*. Finally, we include rules for each information structure we want to recognize using key terms for the respective information and syntactic knowledge for that particular language by allowing optional elements.

Apart from the spelling variation, each information type may occur in several forms. Therefore, if

4. `https://www.facebook.com`
5. `https://foursquare.com`
6. `https://www.instagram.com`
7. `http://english.sina.com`
8. `https://www.waze.com/nl/`
9. We used place names for the Netherlands from GeoNames `http://www.geonames.org/`. The place names 'Brand', 'Gem', 'Wel', 'Een', 'Zuid', 'Noord', 'Oost', 'West' were excluded, since they are highly ambiguous.
10. The temporal expressions list from Hürriyetoğlu *et al.* (2014) was used.

having optional elements does not suffice to model various patterns of occurrence for an information type, we define additional rules for those patterns. The rules in the grammar modules were ordered so that always the left-most longest match applies. This step is the same for rules across information types. The longer a rule, the earlier it is evaluated on a tweet.

We use linguistic structures such as prepositions to restrict ambiguous terms from matching excessively. For instance, the road ID A2 may have irrelevant senses in case it is not preceded by the specific preposition plus article combination *op de* (EN : 'on the').

Below part of one of the grammar modules for location is included as an example. 'Word(alphas)' matches any word that consists of alphanumeric characters with an optional '#' character at the beginning. Matching of the prepositions is represented with the grammar of 'prp'. Finally, 'loc' is defined as the combination of the 'prp', an optional 'sTk' and 'infra' part, and a 'place'. The subpart '~infra + sTk' states the possibility to match any arbitrary element other than an 'infra' element.

inLt = WordStart() + CaselessLiteral('in') + WordEnd()
sTk = WordStart() + Combine(Optional('#') + Word(alphas)) + WordEnd()
prp = (inLt|bijLt|thvLt|voorLt|opLt)
loc = prp + Optional(~infra + sTk + infra) + place

We keep ambiguous domain terms apart from the rest of the terms that are about the same information category. We write more specific rules for them or leave them out. For instance, the term *ongeluk* (EN : 'accident') matches only if it is not preceded by the preposition *per*, because *per ongeluk* (EN :'by accident') is a widely used phrase outside the traffic domain as well. Furthermore, the temporal expression *weer* (EN :'again') is not handled by design because it is ambiguous, the other sense being 'weather'. Weather events that are specified and make use of the word *weer* have to be preceded by an adjective that denotes the kind of weather (e.g. *regenachtig* (EN : 'rainy') or *slecht* (EN : 'bad')).

3.1 Learning Place Names Automatically

Static lexicons are bound to be incomplete. Therefore, we implemented a learning method to extend the lexicon, more specifically with place names. Additional place names or spelling variations of the place names already included in the lexicon were identified (or 'learned') by employing manually crafted linguistic patterns. For instance, the linguistic pattern "tussen <optional infrastructure indicator> <place name 1> en <place name 2>" (EN : 'between <optional infrastructure indicator> <place name P1> and <place name P2>') was used to identify the set of place names in the training data. We used just the P1s as these reliably identified place names, whereas with unknown (multi-token) place names in the position of P2 it is impossible to reliably identify where exactly these end. Note that we discard first names and personal pronouns as possible P1 candidates. The extended lexicon contributes to increasing the coverage. [11]

11. Items such as *Weert-N* in the tweet *Bij Weert is een ongeluk gebeurt op fr #A2 naar Eindhoven, tussen Weert-N. en Budel is the rechterrijstrook dicht. Nu 6 km file.* would otherwise be missed.

4 Data collection and Preprocessing

For development purposes a data set of 85,906 Dutch language tweets was collected over a period of four years (i.e. from January 1, 2011 until March 31, 2015) using the hashtag A2 (#A2), which refers to one of the main highways in the Netherlands, from Twiqs.nl. We eliminated 6,351 tweets that we know are not relevant, for instance tweets about *flitsers* (EN : 'speed cameras'). Moreover, 25 tweets that contain *#A2.0* were eliminated as well. Finally, we excluded all 25,580 tweets that have an indication of being a retweet either in the tweet JSON file or in the lower case form of the text. As a consequence, the final tweet set consists of 57,940 tweets, which were used to identify and manage patterns and to guide us in the rule development phase. This set was not annotated manually, which is a different approach from the typical manner in which a training set is manually annotated to be used for supervised training of the specific information category labels.

An analysis of the remaining tweets revealed that there is still noise and multiple threads of information in the set. The noise mainly arises from the ambiguity of *A2*. Apart from referring to the highway running from Amsterdam to Maastricht, it is also used to refer to a paper size, football teams, school classes, a quality label for real estate, a car type, and reading level among others.

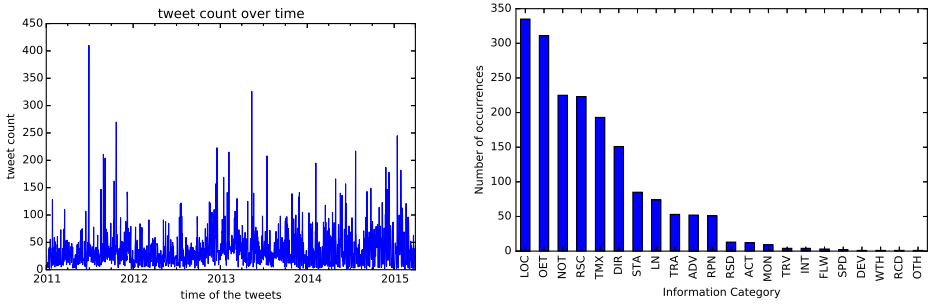
Removing tweets that are not about traffic is not always trivial as domains may overlap. The tweet *sta weer als van ouds vast op #a2 op weg naar training ..* (EN : 'as usual stuck on #a2 on the way to the training') is about traffic on the A2 while at the same time there is reference to the sports domain through the word *training*.

Another source of noise is the tweet language, which - it turns out - is not always Dutch. For example, the tweet *Si el At . de Madrid empata esta noche tampoco estaría tan mal que el #BarÃ§a se ponga #A2.Sería igual de bueno para su visita al Camp Nou !* is obviously in Spanish.

Typical recurring topics in #A2 tweets are traffic jams, warnings about adverse weather or road conditions, and observations about accidents. We observe the following major categories of recurring traffic-related topics :

- **Factual** : an event that happened on a road is explained in detail mainly by the traffic authorities, e.g., *#a2 ri #eindhoven bij afslag pettelaar langzaam rijdend en stilstaand verkeer #vid* (EN : '#a2 richting #eindhoven at pettelaar exit slow-moving and stagnant traffic #vid').
- **Meta** : opinions about traffic situations which are not based on or related to a single event, e.g., *Voedselbanken , dode kinderen Syrie , wereldproblemen maar de #A2 is bij ons issue . .. wat een intellectuele leegheid stralen we toch uit . .. .* (EN : 'Food banks, dead children in Syria , global problems but the #A2 is an issue with us. .. what intellectual emptiness we radiate').
- **User observations** : drivers or passengers commenting on an event, e.g. *Hé politieagentjes op jullie motors ! Voor jullie geldt op #A2 nu ook max 100km . Of doen jullie lampies het soms niet ? #2mate #aso #politie* (EN : 'Hey policemen on your motorcycles ! For you too #A2 is max 100km the limit. Or are your lights not working ? #2mate #aso #politie').
- **User actions** involving a road but not referring to the traffic, e.g., *Ik ga mijn honden uitlaten op de #A2 nu* (EN : 'I am going to walk my dogs on the #A2 now.').

We did not distinguish between these topics in our analysis. We tokenized the tweets and maintained case of the text. The final tweet set contains tweets from 16,555 users. The number of the tweets per user differs immensely : thus there are 12,360 users who each have only one tweet while the other extreme is a single user, @A2Verkeersinfo, that has 13,463 tweets.



(a) Tweet distribution on A2

(b) Information category distribution

FIGURE 1 – Temporal tweet distribution of the training set and information category distribution in the annotated tweets, which is the test set, at the evaluation phase.

The temporal tweet distribution can be observed in Figure 1a. Important events can be recognized as peaks in the tweet count. For instance, the highest peak on June 19, 2011 corresponds to observers' and commentators' tweets about an intense pursuit [12] and its effect on the traffic on the A2.

5 Traffic Domain Modeling

The traffic domain was modeled using human expert knowledge directed at identifying information fields pertaining to the road, traffic, weather conditions, notifications etc. This step enables us to focus on particular parts of the grammar and makes it easier to maintain them. From the tweets we aim to extract, label, and subsequently store in a database the following information reliably.

5.1 Information categories

1. Road ID (RID) : Road identifier ; highways and main roads in the Netherlands are numbered by means of the letter A, E, or N followed by a number.
   — *Ook op de #a2 is het feest URL* (EN (litt.) : 'The party is on the #a2 as well')
2. Road section (RSC) : Stretch of road, frequently indicated by referring to place names and/or infrastructure elements (e.g. intersections or exits).
   — *Kan #A2 van den Bosch naar Boxtel Noord al niet naar 3 baans ? Waarom niet ?* (En : 'Can #A2 from Den Bosch to Boxtel Noord already not be 3 lanes ? Why not ?')
3. Road side (RSD) : Either the left or the right side of the road.
   — *#Flitser #A2 hmp 48,0 li . .. is de trajectcontrole niet effectief genoeg ?* (EN : '#AutomaticSpeedControl #A2 hmp 48,0 left . .. the trajectory control is not effective enough ?')
4. Lane (LN) : Strip of a road, fast lane, right lane etc.
   — *Dichte #Mist op de #A2 in Limburg spitsstrook dicht en adviessnelheid 120km , bijzonder !* (EN : Dense #Fog on the #A2 in Limburg express lane is closed and recommended speed 120km, remarkable !')
5. Road point (RPN) : Detailed or formal specification of a particular place on the road.

12. http://nos.nl/artikel/252005-criminelen-voortvluchtig-na-overval.html

— *USR_a2 : LET OP ! Flitser op de #A2 - hmp 120.8 (Den Bosch -> Eindhoven) #Flitser URL " #ekkersrijt let op !* (EN : 'BE CAREFUL ! Speed camera on #A2 - hmp 120.8 (Den Bosch -> Eindhoven) #Speed camera URL')

6. Road condition (RCD) : State of the road, which can affect the flow of traffic.
   — *Glad in #maastricht en op de #A2 , strooiwagens volop bezig . Blij met mijn ESP en ABS* (EN : 'Slippery in #maastricht and on the #A2 , Spreaders are working hard . Delighted with my ESP and ABS')
7. Direction (DIR) : Geographic directions (e.g. 'east'), or place names that are used in combination with prepositions to indicate direction.
   — *8km file op de #A2 vanuit het zuiden richting Eindhoven , maar de spitsstrook blijft dicht . Goed bezig jongens ! #file USR* (EN : '8km traffic jam on the #A2) from the south towards Eindhoven , but the express lane remains closed . Well done guys ! #TrafficJam USR')
8. Location (LOC) : Specification of a particular place on the road ; less precise than the road point. Places that are not in the Netherlands are out-of-scope.
   — *Kettingbotsing op de #A2 thv Leende* (EN : 'Pileup on the #A2 near Leende')
9. Status (STA) : Information about the status of a road (whether it is open, closed, etc.).
   — *De afrit Echt op de #A2 Maastricht - Eindhoven wordt geblokkeerd door een vrachtwagen met pech .* (EN : 'The exit Echt on #A2 Maastrict - Eindhoven is blocked by a truck with a breakdown .')
10. Traffic (TRA) : Denoting various types of traffic.
    — *Op de #A2 colonnes van touringcars onder politiebegeleiding , met vertraging tot gevolg . Gelukkig ga ik de andere kant op .* (EN : 'On the #A2 convoy of tour buses are under police escort, by causing delay . Fortunately, I'm going the other way.')
11. Flow (FLW) : The status of traffic in terms of flow quality.
    — *4 banen stilstaand verkeer op de #A2 naar t zuiden bij afrit Beesd .* (EN : '4 tracks standstill traffic on the A2 the south at exit Beesd. track')
12. Intensity (INT) : The intensity of the traffic flow.
    — *Wat een rust op de #A2 alweer bijna thuis !* (EN : 'What a peace on #A2 almost home again')
13. Speed (SPD) : Comments about how fast or slow the traffic is moving or what speed limit there is.
    — *Voor degene die naar #utrecht via een #A2 moeten kunnen rekenen op een file . Ter hoogte van #lekbrug rijd men stapvoets* (EN : 'People traveling to #utrecht via #A2 must take into account a traffic jam . Around #Lekburg, the speed is walking pace')
14. Observed event (OET) : Events that can be observed while traveling on the road and that may affect the flow of traffic ; such events include road works, accidents, and other mishaps.
    — *Wegwerkzaamheden veroorzaken nu files op de #A2>Belgische grens voor Maastricht (4 km) en op de #A50>Oss b URL* (EN : 'Roadworks are causing traffic jams now on #A2>Belgian border before Maastricht (4 km) and on #A50>Oss b URL')
15. Weather (WTH) : Mostly (but not necessarily) adverse weather conditions that may have an effect on the flow of traffic.
    — *Strooiploeg , waar ben je ? #A2 #limburg #sneeuw* (EN : 'Sprinkle team , where are you ? #A2 #limburg #snow')
16. Traffic violation (TRV) : Types of traffic violations.
    — *Legaal spookrijdende op #a2 en in file op toerit ! Ze rijden dus. Mooie nacht USR* (EN : 'Legally wrong-way driving on #a2 in traffic jam on ramp ! They drive. Great night USR')

17. Monitoring (MON) : Traffic related controls on the roads by the authorities.
    — *#alcoholcontrole #Breukelen bij viaduct #A2* . (EN : '#alcoholcontrol #Breukelen at crossover #A2 .')
18. Development (DEV) : Updates about the status, road conditions, etc.
    — *File #A2 bij Vught lost snel op* . (EN : 'Traffic jam #A2 near Vught dissolving rapidly .')
19. Activity (ACT) : Actions of the drivers when they face any problem in the traffic.
    — *#verwarde #man zorgt voor #file #a12 Gelukkig op tijd kunnen uitwijken naar #a2* (EN : '#confused man causes #trafficJam #a12 Luckily swerved in time to #a2')
20. Notification (NOT) : News about the traffic and more generally the situation on the road (delays, traffic jams, etc), i.e. the kind of information for which you switch on your radio.
    — *Al de eerste file bij Urmond . #a2* (EN : 'Already the first traffic jam in Urmond . #a2')
21. Advice (ADV) : Explicit mentions of announcements.
    — *Verkeer dat de omleiding volgt i.v.m. het ongevalsonderzoek op #A2 bij #Deil wordt geadviseerd te keren bij Waardenburg (#A2)* (EN : 'Traffic that follows the redirection in relation with the accident investigation near #Deil is suggested to turn at Waardenburg (#A2) ')
22. Time expression (TMX) : Any indication of time, including point in time, time duration, and time frequency.
    — *Of je staat al ruim een uur in de file voor de Belgische grens . Ongeluk met 2 vrachtwagens ? #ongelukA2 #A2* (EN : 'Or you are in a traffic jam already over an hour before the Belgian border. Accident involving two trucks ? #AccidentA2 #A2')
23. Other (OTH) : Traffic related information that is not covered by any of the available categories.

In addition we have (at least) three further bits of information, viz. the time stamp of a tweet, the user who posted it, and the device used to post it. Moreover, occasionally tweets contain the GPS latitude and longitude of the device that sent that post. All these attributes can be used in case the extracted information is not enough to understand what is happening or to disambiguate a relative expression. For instance, based on the number of tweets produced by a user, we can already easily distinguish between professional users and private ones.

6 Results and Discussion

We evaluated our method on a sample of 1,448 Dutch tweets collected between April 1 and April 4 2016, by using the road IDs *A12*, *A28*, *A27*, *A50*, *A7*, and *A58*. We excluded 40 tweets by irrelevant users [13]. Retweets (295) were excluded as well. Finally, we excluded near-duplicate tweets by using a .85 cosine similarity threshold. The final evaluation set contains 728 tweets, in which we observed a few near duplicates.

We created a FoLiA document (van Gompel & Reynaert, 2013) from batches of 250 tweets and used the web-based annotation tool FLAT [14] to manually annotate the tweets [15]. Annotated tweets contain the information categories displayed in Figure 1b. The figure does not contain the road ID category, since every tweet contains a road ID due to the tweet collection method used. The near absence of the Other (OTH) category indicates that the information categories are fairly complete for this domain.

13. These users are identified manually in the training set.
14. `https://github.com/proycon/flat`
15. A single trained person performed this task.

720 tweets contain both manually annotated and automatically detected information [16]. The total number of manually annotated and automatically detected information units are 2,699 and 2,400 respectively. On the one hand, 285 of the manual annotations were not identified by the rules. On the other hand 93 information units that were detected by the rules did not match any unit in the manual annotations. The automatic rule-based method mostly failed to capture locations that are not preceded by a preposition (such locations were ignored by design). Moreover, some tokens that are not in the scope of any relevant information category were mistakenly identified as temporal expressions.

The automatic method detected exactly the same information with the annotations in 1,245 cases. In 79 cases the matched tokens were the same, but the information category did not match. Most of the errors were caused by the confusion between direction and road section categories. Annotated and automatically detected units overlapped 542 times with the same information category. However, in 73 cases the information category was different. In the overlapping cases 452 of the time the automatic method detected longer phrases, which mostly cover the preceding prepositions and articles (included intentionally for reasons of maintaining control). In 160 cases the manual annotations were applied to longer phrases [17].

Evaluating an information extraction method, especially when the frequency of the information categories are unbalanced, with only the scores has various drawbacks (Esuli & Sebastiani, 2010). However, we still report these performance scores to provide a rough summary of the performance. The precision of the correct information category detection is 51% and 74% for the exact and overlapping token matches, respectively. The recall is 46% and 66% in the same scope. These numbers are based on performance on individual tweets. Having multiple tweets about a single traffic event will increase the chance of detecting these smaller events.

7 Conclusion and Future Work

We presented our study on collecting traffic related tweets, analyzing and preprocessing these tweets, and extracting detailed information from them that might be used in order to increase traffic safety. Our method performs reasonably well especially on the more frequent information categories.

Our next steps will be directed at evaluating the method on rare information categories and on a set of tweets about a particular event, at extending the coverage, e.g. with automatic coverage extension methods such as proposed by Buchholz & Van den Bosch (2000), detecting and handling the irrelevant senses of the domain terms, learning domain terms from the tweets in addition to place names, and at extracting information from multi-token lexical items.

Acknowledgment

This research was funded by the Dutch national research programme COMMIT and is supported by Statistics Netherlands (CBS). We thank Servan Bulut and Mustafa Erkan Başar for performing the detailed annotation and developing the demo website respectively.

16. All tweets contain a token matching a road ID due to the collection method, but they did not denote road IDs in eight of the cases

17. The sum of the unit numbers does not match the total due to 1-to-many matches.

Références

ANANTHARAM P., BARNAGHI P., THIRUNARAYAN K. & SHETH A. (2014). Extracting city traffic events from social streams. *ACM Transactions on Intelligent Systems and Technology*, **9**(4).

AXEL SCHULZ, PETAR RISTOSKI & HEIKO PAULHEIM (2013). I See a Car Crash : Real-time Detection of Small Scale Incidents in Microblogs. In *The Semantic Web : ESWC 2013 Satellite Events*, p. 22–33. Springer Berlin Heidelberg.

BUCHHOLZ S. & VAN DEN BOSCH A. (2000). Integrating seed names and n-grams for a named entity list and classifier. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, p. 1215–1221, Athens, Greece.

CUI J., FU R., DONG C. & ZHANG Z. (2014). Extraction of traffic information from social media interactions : Methods and experiments. In *IEEE 17th International Conference on Intelligent Transportation Systems (ITSC)*, p. 1549–1554 : IEEE.

D'ANDREA E., DUCANGE P., LAZZERINI B. & MARCELLONI F. (2015). Real-Time Detection of Traffic from Twitter Stream Analysis. *IEEE Transactions on Intelligent Transportation Systems*, **16**(4), 2269–2283.

ENDARNOTO S. K., PRADIPTA S., NUGROHO A. S. & PURNAMA J. (2011). Traffic condition information extraction & visualization from social media twitter for android mobile application. In *Electrical Engineering and Informatics (ICEEI), 2011 International Conference on*, p. 1–4 : IEEE.

ESULI A. & SEBASTIANI F. (2010). Evaluating Information Extraction. *Multilingual and Multimodal Information Access Evaluation*, p. 100–111.

GAL-TZUR A., GRANT-MULLER S. M., KUFLIK T., MINKOV E., SHOOR I. & NOCERA S. (2015). Enhancing transport data collection through social media sources : methods, challenges and opportunities for textual data. *IET Intelligent Transport Systems*, **9**(4), 407–417.

HE J., SHEN W., DIVAKARUNI P., WYNTER L. & LAWRENCE R. (2013). Improving traffic prediction with tweet semantics. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, p. 1387–1393 : AAAI Press.

HÜRRIYETOĞLU A., OOSTDIJK N. & VAN DEN BOSCH A. (2014). Estimating time to event from tweets using temporal expressions. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, p. 8–16, Gothenburg, Sweden : Association for Computational Linguistics.

HUZITA E. H., DE SOUZA T. G. & KABUKI Y. H. (2012). A system to capture and generation of traffic information from posted messages on social networks. In *Collaborative Systems (SBSC), 2012 Brazilian Symposium on*, p. 174–180 : IEEE.

LEDUC G. (2008). Road traffic data : Collection methods and applications. *Working Papers on Energy, Transport and Climate Change*, p. 1–55.

MCGUIRE P. (2007). *Getting started with pyparsing*. "O'Reilly Media, Inc.".

TOSTES A. I. J., SILVA T. H., DUARTE-FIGUEIREDO F. & LOUREIRO A. A. F. (2014). Studying traffic conditions by analyzing foursquare and instagram data. *ACM PE-WASUN'14*, p. 17–24.

VAN GOMPEL M. & REYNAERT M. (2013). Folia : A practical xml format for linguistic annotation - a descriptive and comparative study. *Computational Linguistics in the Netherlands*, **3**, 63–81.

WANICHAYAPONG NAPONG, PATTARA-ATIKOM WASAN P. & RATCHATA (2014). Road Traffic Question Answering System Using Ontology. In *Semantic Technology*, volume 8943, p. 422–427. Springer International Publishing.

WIBISONO A., SINA I., IHSANNUDDIN M. A., HAFIZH A., HARDJONO B., NURHADIYATNA A., JATMIKO W. & MURSANTO P. (2012). Traffic Intelligent System Architecture Based on Social Media Information. In *Advanced Computer Science and Information Systems (ICACSIS), 2012 International Conference on*, p. 25–30.