Information extraction from social media: A linguistically motivated approach

Nelleke Oostdijk[1], Ali Hürriyetoğlu[1,2], Marco Puts[2], Piet Daas[2], Antal van den Bosch[1]

[1]Centre for Language Studies, Radboud University, Nijmegen
[2]Centraal Bureau voor de Statistiek

LANGUAGE MACHINES

CLST | Centre for Language and Speech Technology
Radboud University

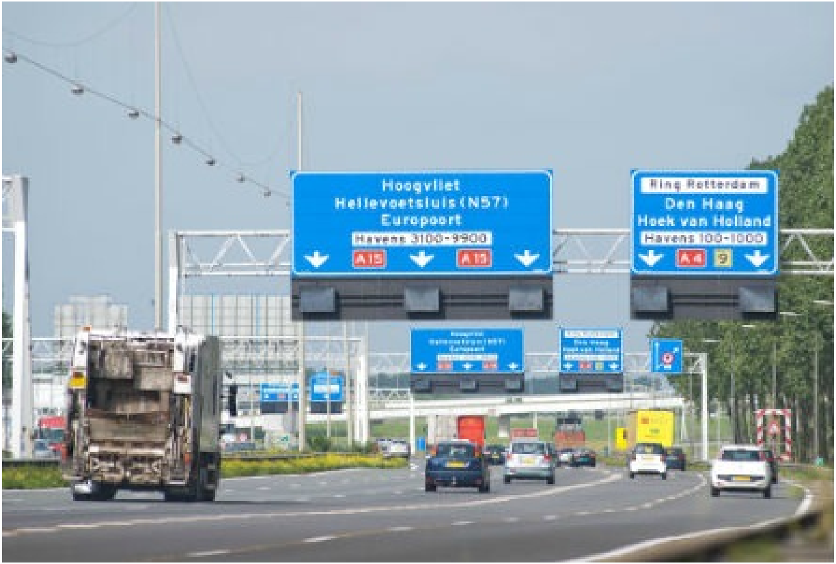
Traffic information extraction

Social media data are used to obtain information

- about the flow of traffic on Dutch main roads (heavy traffic, traffic jams, accidents, adverse weather conditions, blocked roads, etc.)
- about how social media users are coping with or reacting to a particular situation
- in addition to data from other sources (e.g. sensors, weather forecasts, reports filed by the authorities)

Twitter

- Freely accessible, real-time, widely used, rich source of information; tweets come with time stamps and many also with geo-location


Hoogvliet
Hellevoetsluis (N57)
Europoort
Havens 3100-9900
A15
A15
Ring Rotterdam
Den Haag
Hoek van Holland
Havens 100-1000
A4
9

Extracted traffic information

1. **Activity**
2. Advice
3. Development
4. Direction
5. Flow
6. Intensity
7. Lane
8. **Location**
9. Monitoring
10. Notification
11. **Observed event**
12. Road condition
13. Road ID
14. **Road point**
15. Road section
16. Road side
17. Speed
18. Status
19. **Time expression**
20. Traffic
21. Traffic violation
22. Weather

Use cases

Up-to-date traffic information can improve life for all of us!

Use cases

Informed drivers will

- Be less likely to be caught up in traffic jams (e.g. they can take an alternative route)
- Experience a safer journey (e.g. because they have been forewarned about icy roads)
- Be able to plan their journey better (e.g. decide to leave home earlier)
- Be encouraged to share their observations

Authorities will be able to

- Quickly respond to traffic situations
- Use information about new events
- Get a better insight into the status of various traffic events
- Understand the risks for traffic on the roads better

Use cases

Extracted information can be used in combination with

- Data from traffic sensors
- Information about the existing or expected weather conditions
- Information about ongoing or planned road works
- Information about ongoing or planned events
- Information as regards traffic route planning

Information extraction method

We process the flexible language use found on social media by a

- Rule-based, hierarchical, formal information extraction methodology
- Partial match approach for traffic domain terms, place names, and time expressions
- Pattern-based token representation that allows the partial coverage of tokens, while tokens may start with a hash (‘#’) or end with a period (‘.’).

This methodology improves the flexibility of the lexical and syntactic structures that can be covered.

Information extraction method

Location Grammar

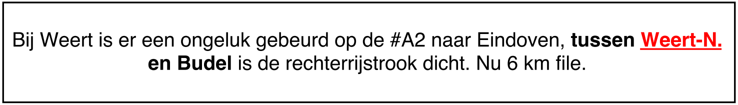
```
inLt = WS() + CaselessLiteral(‘in’) + WE()
sTk = WS() + Combine(Optional(‘#’) + Word(alphas)) + WE()
prp = (inLt|bijLt|thvLt|voorLt|opLt)
loc = prp + Optional(~infra + sTk + infra) + place
```

Flexible Token

```
optLt = Optional(Literal("#"))
road_tok = oneOf(["weg","route","straat","laan"], caseless=True)
r_cntxt = SkipTo(road_tok, include=True, failOn=White())
token_roadx = Combine(optLit + Word(alphas, exact=1) + r_cntxt)
token_roadx.setParseAction(validateString)
```

Learning place names

- Place names are detected in slots in linguistic patterns:
  - “tussen <optional infrastructure indicator> <place name P1> en <place name 2>” (EN: between <optional infrastructure indicator> <place name P1> and <place name P2>)



Bij Weert is er een ongeluk gebeurd op de #A2 naar Eindoven, **tussen Weert-N. en Budel** is de rechterrijstrook dicht. Nu 6 km file.

Detecting new place names

  - We filter out person names with exception lists for a cleaner list of place names.

Development data

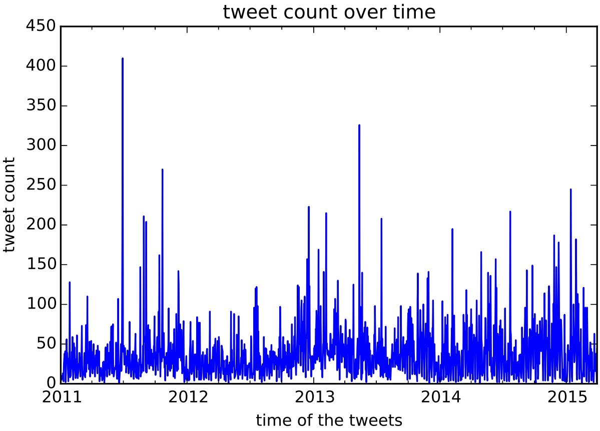
- 85,906 Dutch language tweets were collected over a period of 4 years (i.e. from January 1, 2011 until March 31, 2015) using the hashtag A2 (#A2).
- We removed 6,351 tweets from users that we know are not relevant, for instance users dedicated to tweet about "flitsers" 'detectors of speed violations'.
- Moreover, 25 tweets that contain "#A2.0" were removed as well.
- Finally, we excluded all retweets (25,580).
- As a consequence the final tweet set consists of 57,940 tweets.

Remaining noise in the data

Although we observed the following additional meanings of the key term A2, we did not exclude the tweets that has this meaning.

1. some different meaning in a foreign language
2. paper size
3. football team
4. school classes
5. a quality label for real estate
6. a car type
7. reading level

Our grammars does not take into account any meaning other than the road ID for the key term A2.



A2 tweet distribution over time

Main threads of information

- **Factual**: an event that happened on a road is explained in detail mainly by the traffic authorities.
- **Meta**: opinions about traffic situations which are not based on or related to a single event.
- **User observations**: drivers or passengers commenting on an event
- **User actions** involving a road but not referring to the traffic.

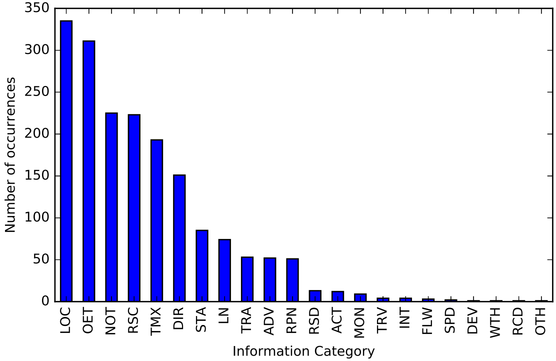
Information extraction examples

- 8km file op de #A2 vanuit het zuiden richting Eindhoven , maar de spitsstrook blijft dicht . Goed bezig jongens ! #file @vanAnaarBeter
- Lekker in de file op de #A2 , tussen knp. Deil en Nieuwegein-Zuid 16 km stilstaand verkeer (vertraging : meer dan 30 min) aldus de @vid
- Strooiploeg , waar ben je ? #A2 #limburg #sneeuw

Evaluation Data

- We collected 1,448 Dutch tweets between April 1 and 4 using the road IDs *A12*, *A28*, *A27*, *A50*, *A7*, and *A58*.
- We excluded 40 tweets based on the information we observed in the development data, e.g., users (40 tweets), retweets (295 tweets).
- We kept only one of the near-duplicate tweets, tweets that have above .85 cosine similarity are accepted as near-duplicates.
- The remaining set contains 728 tweets.

Manual annotation



Information category distribution in the annotations that were performed by one annotator by using the FLAT (`https://github.com/proycon/flat`) platform.

Results on individual tweets

- The total number of manually annotated and automatically detected information units are **2,699** and **2,400** respectively.
- **285** of the manual annotations and **93** of the automatically detected information units were not observed on the other side of the comparison.
- 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.

Results on individual tweets

- Annotated and automatically detected units overlapped **542** and **73** cases with the same and different information category respectively.
- In the overlapping cases **452** of the time the automatic method detected longer phrases, which mostly cover the preceding prepositions and articles.
- In **160 cases** the manual annotations were applied to longer phrases compare to automatic method detects.
- 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.

Discussion

- Most of the errors were caused by the confusion between direction and road section categories.
- The automatic method mostly failed to capture locations that are not preceded by a preposition (such locations were ignored by design).
- Some tokens that are not in the scope of any relevant information category were mistakenly identified as temporal expressions.
- Having multiple tweets about a single traffic event will increase the chance of detecting these smaller events.

Next steps

- Create a filter to be able to use a more general tweet stream.
- Increase the lexical coverage: plural forms of (noun) tokens, verb tenses (third person), 010 etc. for place names.
- Include names of rivers (for inclusion in rules that make use of words referring to parts of the roads infrastructure, the bridge across the river X, the tunnel under the river Y), e.g.
- Identify discourse units.
- Do more online learning.
- Add different makes of cars and the types: Audi, Mercedes, Opel, etc.
- Evaluate the system by comparing amount of information captured with a standard traffic information database.

Acknowledgements

This research was funded by the Dutch national COMMIT programme and is supported by CBS.

Centraal Bureau voor de Statistiek

COMMIT/

#A2 tweets were retrieved from http://twiqs.nl.

Pyparsing library played a key role in development of the information extraction method.

Tweets are stored in MongoDB.

Thanks for listening. Any questions or comments?

Please find more information on:

- http://sinfex.science.ru.nl
- https://bitbucket.org/hurrial/sinfex

Contact

Nelleke Oostdijk, *n.oostdijk@let.ru.nl*
Ali Hürriyetoğlu, *a.hurriyetoglu@let.ru.nl* (@hurrial)
Marco Puts, *m.puts@cbs.nl* (@MarcoPuts)
Piet Daas, *pjh.daas@cbs.nl* (@pietdaas)
Antal van den Bosch, *a.vandenbosch@let.ru.nl* (@antalvdb)