Evaluation of Named Entity Recognition in Dutch online criminal complaints

Marijn Schraagen
Floris Bex
Matthieu Brinkhuis

Utrecht University

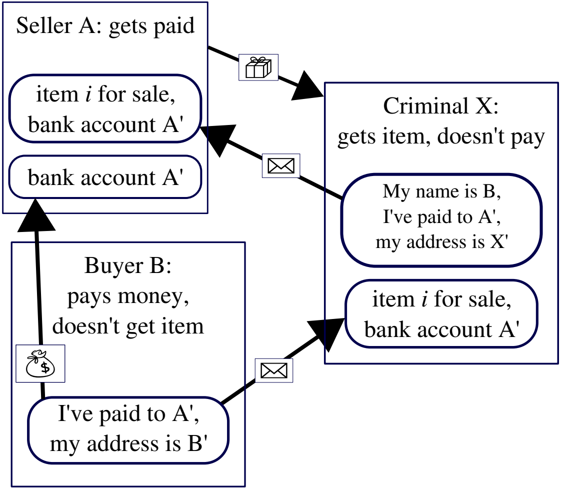
June 12, 2017

Universiteit Utrecht

Internet fraud

- Online trade is widespread
- Transactions involving commercial parties or private sellers
- Several types of fraud are addressed by the Dutch Police
  - As buyer: product payed but not delivered
  - As seller: product delivered but no payment received
  - Triangle fraud

Triangle fraud



Submitting a crime report on politie.nl

Betreft het een handelssite? *

Advertentietitel

Advertentienummer

Uw gebruikersnaam handelssite *

Gebruikersnaam wederpartij handelssite

Omschrijving conflict *

Heeft u nog opmerkingen of andere relevante informatie met betrekking tot uw melding? Is u bijvoorbeeld iets opgevallen aan de adverteerder, zijn advertenties of personalia? U kunt dit hier invullen

Relation extraction

- Every year around 50k crime reports are submitted
- Free text entry contains important entities and relations
- I bought item A on webshop-X-deals.com. I called webshop-X, who informed me that webshop-X-deals.com are imposters.
- Relation extraction: `bought(submitter,item_A,webshop-X-deals.com)`
- Need for automatic processing
- *Note:* for privacy reasons all examples are anonymised or constructed

Named entity recognition

- `bought(submitter,item_A,webshop-X-deals.com)`
- Desired: accurate named entity recognition
- Data not optimal: real-world, submitted on internet form, many spelling and grammar errors
- **How accurate are current tools on this type of data?**
- This evaluation: Dutch NLP suite *Frog*
  - Statistical machine learning
  - 80% accuracy on test corpus: newspapers, magazines, Wikipedia pages, brochures, autocues, ...

Entity type

- Frog uses six standard *enamex* entity types
  - Location
  - Person
  - Organisation
  - Event
  - Product
  - Miscellaneous
- Metonymic type
  - **Spain** has won the world cup.
  - *location* metonymically used as *organisation*

Reference set

- Dataset with 64k crime reports
  - typically 1-5 sentences (85 tokens on average)
- Create gold standard reference set using two expert annotators, indicating scope, type, and metonymic type
  - 250 documents, 23k tokens, 1191 named entity tokens

> Purchase of **Iphone 5s**$_{\text{PRODUCT}}$ on **marktplaats**$_{\text{ORG.LOCATION}}$. 250 euro transferred to the account of **John Doe**$_{\text{PERSON}}$ trusting that he would send the **iphone**$_{\text{PRODUCT}}$ by registered mail. The next day I received a message from **marktplaats**$_{\text{ORGANISATION}}$ that the account of **John Doe**$_{\text{PERSON}}$ is fraudulent. I have therefore transferred money to an account of a swindler named **John Doe**$_{\text{PERSON}}$.

Evaluation setup

- Perform NER using Frog
- Compare results on precision and recall
- Classification can be partially correct
  - Type errors
  - Scope errors

Scope errors

- NE's can be single-word *(John)* or multi-word *(John Doe)*
- Recognition can be incomplete
  - I bought an Iphone 5s **Gold** with invoice
  - *Correct:* I bought an **Iphone 5s Gold** with invoice
- Recognition can be overcomplete
  - He lives in **Amsterdam The** next day I called
  - *Correct:* He lives in **Amsterdam** The next day I called

Evaluation results

	category	precision	recall	F-score
1.	entity detected	0.83	0.61	0.71
2.	scope correct	0.63	0.54	0.58
3.	type or metonymic type correct	0.47	0.47	0.47
4.	type correct	0.43	0.45	0.44
5.	scope and type correct	0.35	0.40	0.38

Conclusion: the Frog NER-module does not provide adequate performance for unedited non-professional text

Results per entity type

	event	location	misc	organisation	person	product
correct	0.0	0.73	0.09	0.16	0.57	0.00
scope error	0.3	0.14	0.16	0.06	0.23	0.19
scope correct, type error	0.4	0.07	0.14	0.35	0.08	0.25
not recognised	0.3	0.06	0.60	0.43	0.12	0.56

Observation: recognition of locations and persons is relatively accurate

Error examples

error type	*example sentence*
no error	and **John Doe**$_{\text{PERSON}}$ didn't respond to my messages
wrong type	On **Marktplaats**$_{\text{PERSON}}$ **I** bought shoes
too narrow	**I** transferred money to **NL01** ABCD 1234 5678 90
too wide	He lives in **Amsterdam The** next day **I** called
incorrect	Very **bad reviews**.
unclear	talked on WhatsApp: [01/01 10:00] **See You**: thanks

Error categories

property	amount	proportion
brand or company name	156	0.21
capitalisation incorrect	94	0.13
(alpha)numerical code	39	0.05
punctuation incorrect	35	0.05
partial or full url	32	0.04
bank account number	27	0.04
start of sentence capital	26	0.04
e-mail address	24	0.03
bank country code as location	15	0.02
abbreviation	14	0.02

Possibilities for improvement

- Extend gazetteer
  - 151/156 misclassified brand names: [Marktplaats, Whatsapp, Facebook, Paypal, Google]
- Missing entity categories
  - IBAN, e-mail addresses, urls, miscellaneous codes
  - Pattern matching approach
  - Alternatively: add to training data
- Other errors are more difficult to address
  - Punctuation, capitalisation

Linguistic objectives vs. application domain

- Research objective: relation extraction
- Police use case: filter *relevant* pieces of information
- Objective of NLP tools: find entities that are named
- Definition and scope varies
  - 70 page annotation guideline necessary
- Objectives do not necessarily coincide

Linguistic objectives vs. application domain

- Objectives of NLP tools and current research do not necessarily coincide
- Influences evaluation setup
- Influences evaluation interpretation
- Influences approach to improve NER

NER improvements

- Insufficient amount of data for full retraining of NER algorithm
    - Obtaining domain/genre-specific data may be possible
    - Cf. 60k annotated entities in original training corpus
- Suggested approach: pre- or post-processing of text
- Current evaluation: focus on organisation, product
- Future work: attempt relation extraction in order to discover which errors are most problematic for end application

Questions?

M.P.Schraagen@uu.nl