De-identification in Natural Language Processing

Veronika Vincze and Richárd Farkas
University of Szeged, Department of Informatics
Introduction

• **Natural language processing (NLP):** interdisciplinary field between linguistics and artificial intelligence

• **To understand human language with automatic methods**

• **Information extraction:** collecting relevant information from unstructured texts to gather new knowledge
Databases in NLP

• Supervised learning: from annotated texts, machine learning algorithms generalize patterns which are later applied to unannotated texts

• Large annotated databases are needed for this
Data protection issues

- Annotated texts should be cleared from sensitive and personal data before publication
- Here we examine de-identification in 3 areas:
  - Clinical NLP
  - NLP for social media
  - Information extraction from CVs
Named entity recognition

- Named entities (NEs): names or identifiers that uniquely refer to one entity in the world:
  - Person names (*John*)
  - Organizations (*United Nations*)
  - Locations (*Opatija*)
  - Product names (*Volvo*)
  - Identifiers (*john@gmail.com*) etc.

- Named entity recognition: automatic identification and classification of NEs
- Information extraction usually seeks to find relations between NEs
- Categories to be de-identified are usually NEs
State-of-the art NER

- Challenges and shared tasks for NER:
  - MUC-7 (Chinchor 1998)
- Typical approach: sequence labeling (CRF)
- Best performing systems: F-scores of 0.85-0.89
- NER techniques also applicable in de-identification
De-identification in clinical NLP
De-identification in medical NLP

• Clinical documents contain personal health information (PHI)

• HIPAA (Health Information Portability and Accountability Act) requires to remove:
  – Names of patients and relatives
  – Names of doctors
  – IDs
  – Telephone, fax, pager numbers
  – Hospital names
  – Geographic locations
  – Dates
Anonymization

- Sensitive data is replaced by another one of the same category (e.g. name for a name)
- The text remains readable
- Even if anonymization is not perfect, the user will not know which data is original and which is replaced
State-of-the-art de-identification methods

- Rule-based systems or statistical methods (Douglass et al. 2005, Sibanda – Uzuner 2006)
- NER system adapted to the clinical domain: 0.967 F-score (Szarvas et al. 2007)
De-identification in NLP for social media
Social media

• Communication through social media gains importance
• User-generated content may tell a lot about the user him/herself
• Personal data can be easily gathered from user profiles
Databases of social media texts

- Some attempts to publish anonymized data from social media
  - Sentiment analysis from tweets (Wilson et al., 2013)
  - Shared task on computational personality recognition (Celli et al. 2013) based on Facebook status updates deprived from personal data
Personalization

• Goal: to provide each user with a personalized offer

• The profiling system makes suggestions on the basis of (1) the user’s earlier preferences or (2) generalizations from the choices of users having similar characteristics to the user in question

• Social media is a gold mine for gathering reviews and opinions

WHAT A DISAPPOINTMENT!!!!! I was a huwe fan of Blackberry Bold 9700, but it was getting old and decided to buy the q5. I cannot express how disappointed I am!

--- a 40 year old woman did not like Blackberry Q5
Privacy issues

• Some of the personal data should be preserved since otherwise no useful information would remain

• Templates should be applied

  A <gender> between <age group> did not like the <part> of <mobile>

  A male between 15 and 18 did not like the screen of iPhone4

• The user’s characteristics can be compared to those stored in the database in order to give personalized offers:

  What phones are mostly liked by teenage girls?
Categories for sensitive data treatment

- **Data to be removed / replaced**: person names, addresses, phone numbers, e-mail addresses and other identifiers that uniquely refer to the person (Skype, Twitter IDs etc.)…

- **Data to be merged into groups**: the age of the person (transformed into age groups), the city where the user lives (big city / town / village)…

- **Data to be preserved**: names of products or organizations (for the personalization case)
De-identification for information extraction from CVs
De-identification in IE from CVs

• In order to facilitate candidate selection for a given position, automatic methods are needed to collect relevant information from CVs

  – **Personal data**: name, date of birth, address, phone number, e-mail, marital status
  – **Education**: name of schools, degrees
  – **Employment**: companies the candidate previously worked for, earlier jobs or positions
  – **Language skills**
  – **Competencies**
  – **Hobbies and interests**
CV databases (?)

- For IE systems to be trained, databases of CVs should be available
- Due to data privacy issues, we are not aware of public CV databases
- Sensitive data should be anonymized before publication
Categories for sensitive data treatment

- **Data to be replaced**: names, addresses, phone numbers and emails
- **Data to be preserved**: data items concerning education (degrees, college / university) and data items concerning employment (positions, earlier employers)
- **Data to be deleted**: age, gender and marital status (to assure equal rights)
Conclusions

• De-identification issues in NLP
• For the publication of datasets, de-identification of sensitive data is essential
• Clinical NLP: NER tools are used for anonymization
• NLP for social media: there are some anonymized datasets
• NLP for CVs: no such dataset
• Future work: implement tools for de-identification of CVs or texts from social media
References


