Named Entity Recognition

A brief introduction / problems / solutions

Alexander Kuscher & Christoph Thiele

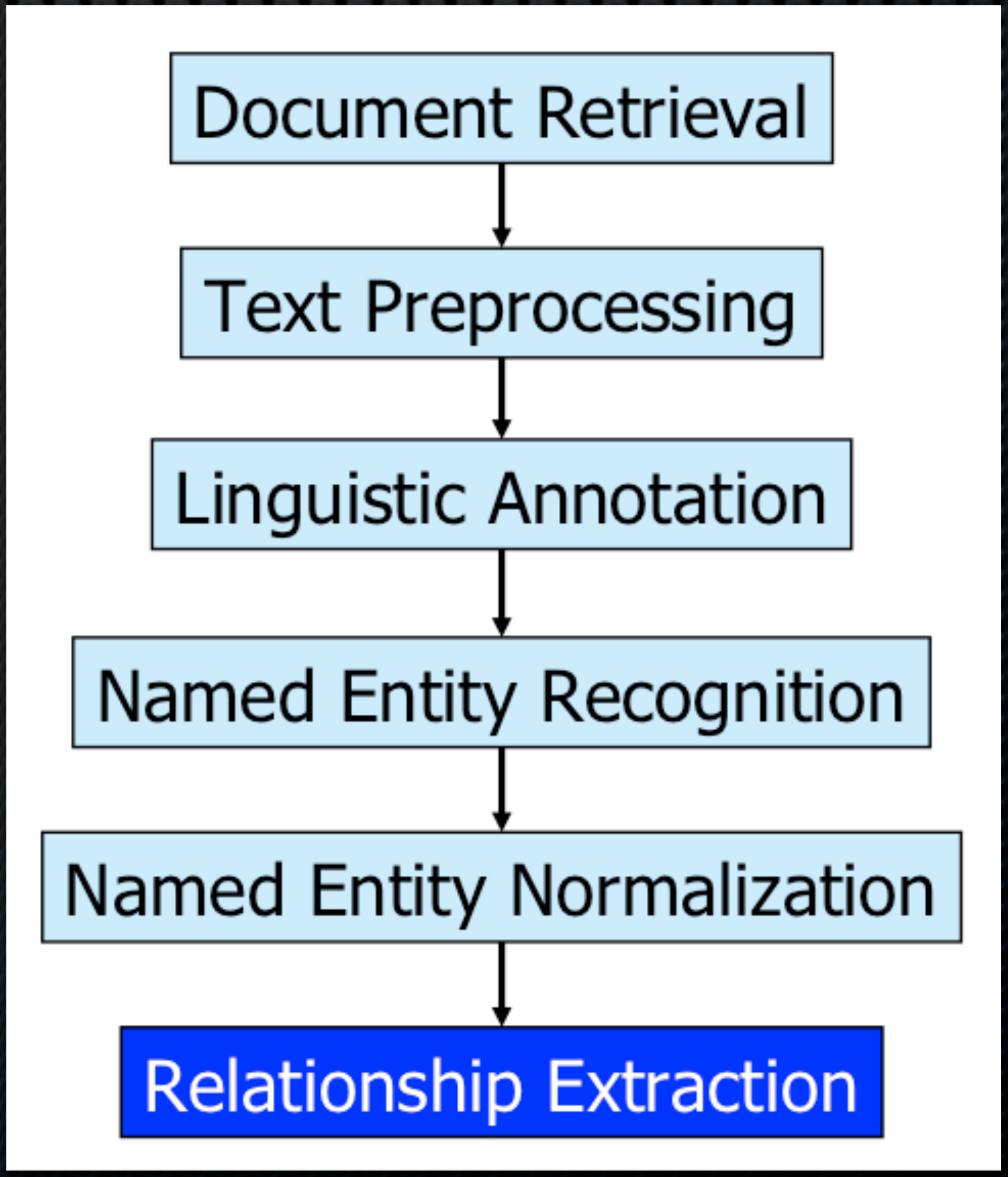
Contents

- Named Entity Recognition - short introduction
- Problems of the domain
- Different Approaches
    - List based approach
    - Regex - divide and conquer
    - Progressive Techniques
    - Association Measurement
    - Insight into our solution
- Do and don‘t do

Short introduction

- **Named entity recognition** (NER) is a subtask of information extraction that seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

Wikipedia



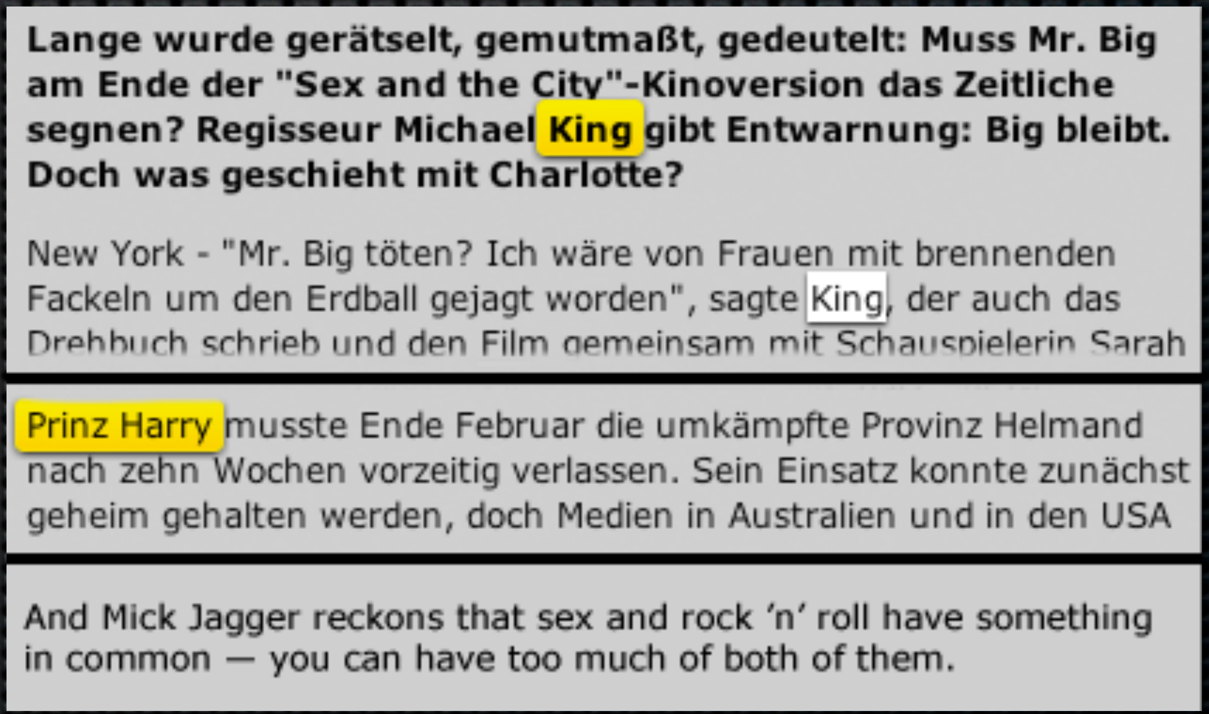
Quelle: Prof. Ulf Leser

Vocabulary

- Precision
  - How much of your entities are correctly identified?
- Recall
  - Number of found entities vs. number of existing ones
- F-measure
  - Harmonic mean of precision and recall

Problems of our domain

- German language
  - all nouns are written with capital letters
- Pseudonyms
- Names and normal nouns overlap
- Titles and peerages
- Forenames w/o surnames
- English texts are sometimes easier



Lange wurde gerätselt, gemutmaßt, gedeutelt: Muss Mr. Big am Ende der "Sex and the City"-Kinoversion das Zeitliche segnen? Regisseur Michael King gibt Entwarnung: Big bleibt. Doch was geschieht mit Charlotte?

New York - "Mr. Big töten? Ich wäre von Frauen mit brennenden Fackeln um den Erdball gejagt worden", sagte King, der auch das Drehbuch schrieb und den Film gemeinsam mit Schauspielerin Sarah

Prinz Harry musste Ende Februar die umkämpfte Provinz Helmand nach zehn Wochen vorzeitig verlassen. Sein Einsatz konnte zunächst geheim gehalten werden, doch Medien in Australien und in den USA

And Mick Jagger reckons that sex and rock 'n' roll have something in common — you can have too much of both of them.

List based approach

- Extract a list from e.g. wikipedia to get possible entities
- Apply this list to your data

- Pros
  - Yields high precision result
  - Low complexity
- Cons
  - You‘ll miss names that are not in your list
  - Low recall
- Extension
  - More sources for entities
  - Weight different lists

Regex - divide and conquer

- One regular expression is not likely to cover all entities
- Different regular expressions for different criteria of the entities in your domain
  - Can cause overlapping results or duplicates
  - Combining those results is sometimes difficult
  - Possible solution might be an internal evaluation system

Progressive Techniques

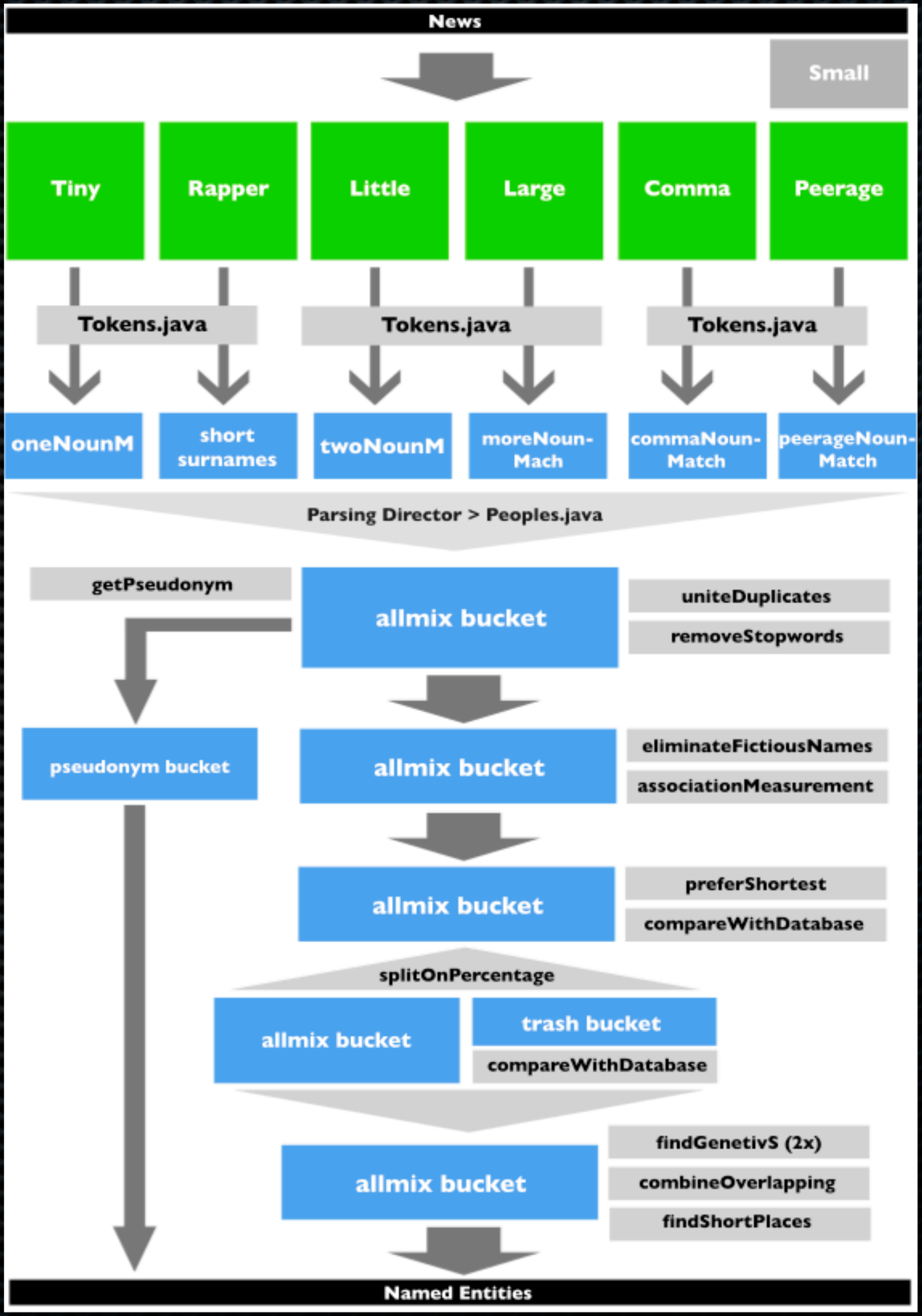
- Take your results as a dictionary for a second / third / fourth ... run
  - Instead of finding only one occurence of an entity you‘ll find nearly every occurence
  - Add special runs for special cases (genitiv)
  - Be careful because if there is waste your source you‘ll get even more

Association Measurement

- Multiword expressions can be used to identify wordgroups that seem to belong together
  - Multiword expression = two or more lexems
    - lexem = ran, run, runs
  - e.g. noungroups can be identified as names
    - „Hans Müller hat [...]. Auch ist Hans Müllers [...]“

Our solution

- Several Regex
- Own data structure
- Duplicate elemination (due to different Regex)
- Stopword list
- Several Dictionaries
- Postprocessing
  - Genetiv
  - Overlapping
  - Too long entities



Do and don‘t do

- Do
    - Adaptive approaches are very effective
    - Use knowledge from SWA / SWE
        - Modular Design
        - Interfaces
    - Prototyping
    - Fail often and fail early!
    - Mix different approaches (list based, AM, PT)
- Don‘t do
    - Try putting all logic into one regular expression
    - Only list-based approach
    - Focus only on precision or recall
    - ...

?