NewsReader: Automatically extracting Events, Entities and Perspectives from Newspapers

Marieke van Erp
marieke.van.erp@vu.nl
http://mariekevanerp.com

• Consortium: Vrije Universiteit Amsterdam (NL), The University of The Basque Country (ES), Fondazione Bruno Kessler (IT), LexisNexis (NL), ScraperWiki (now “The Sensible Code Company”, UK) & SynerScope (NL)

• **Read** massive streams of news from many different sources

• **Record** the changes in the world as they are told in the sources in 4 languages: English, Dutch, Spanish and Italian.

• **What** happened, **where** and **when**, **who** was involved.

• From unstructured **Text** to structured **RDF** (through a happy marriage between Computational Linguistics and Semantic Web researchers).

• Who made what statement, where do sources agree and disagree, what is their emotion or judgement: **provenance**

http://www.newsreader-project-eu
17/06/2013

Porsche family buys back 10pc stake from Qatar

Descendants of the German car pioneer Ferdinand Porsche have bought back a 10pc stake in the company that bears the family name from Qatar Holding, the investment arm of the Gulf State’s sovereign wealth fund.

All of the common shares in Porsche Automobil Holding SE are now held by the Porsche-Piech family. Descendants of the eng-

Qatar Holding sells 10pc stake in Porsche to founding families

Qatar Holding, the investment arm of the Gulf state’s sovereign wealth fund, has sold its 10pc stake in Porsche SE to the luxury carmaker’s family shareholders, four years after it first invested in the firm.

Qatar Holding, which owns stakes in some of the world’s largest companies, said it sold the common shares in the automaker to the Porsche and Piech families. It did not disclose the value of the transaction.
Natural Language Processing Pipeline
NLP Annotation Format

• Stand-off XML

• Based on KAF, TAF, LAF and uses URIs (from RDF)

• NAF-FoLiA converters are in progress

• Each annotation receives a new layer
Semantic Annotation

- Named Entity Recognition & Linking
- From words to concepts
- Semantic Role Labelling
- Recognising Temporal Expressions & Relations
- Wikification
Named Entity Recognition & Linking


<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>NewsReader</td>
<td>92.20</td>
<td>90.19</td>
<td>91.18</td>
</tr>
<tr>
<td>(ixa-pipe-nerc)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stanford NER</td>
<td>89.37</td>
<td>87.95</td>
<td>88.65</td>
</tr>
<tr>
<td>Ratinov et al. (2009)</td>
<td>-</td>
<td>-</td>
<td>90.57</td>
</tr>
<tr>
<td>Passos et al. (2014)</td>
<td>-</td>
<td>-</td>
<td>90.90</td>
</tr>
</tbody>
</table>

NERC CoNLL 2003 testb results.
Named Entities in NAF

<entities>
  <entity id="e1" type="EVE">
    <references/>
    <span>
      <!--Honda Jazz-->
      <target id="t_22"/>
      <target id="t_23"/>
    </span>
  </references>
  <externalReferences>
    <externalRef confidence="0.9999979"
      reference="http://nl.dbpedia.org/resource/Honda_Jazz" reftype="nl" resource="dbpedia-nl"
      source="spotlight_v1"/>
  </externalReferences>
</entity>

<entity id="e2" type="MISC">
  <references/>
  <span>
    <!--Belaastingdienst-->
    <target id="t_193"/>
  </span>
</entity>

<externalReferences>
  <externalRef confidence="1.0"
    reference="http://nl.dbpedia.org/resource/Belaastingdienst" reftype="nl" resource="dbpedia-nl"
    source="spotlight_v1"/>
</externalReferences>
</entities>
Why link to a resource such as DBpedia?

• It allows you to query for fine-grained entity types: give me all politicians in the dataset, give me all football players

• Plus: the background knowledge provides additional filters: give me all politicians born after 1900 in the dataset

• Caveat: the background knowledge is not complete
Analyzing the encyclopaedic novel: ‘The Discovery of Heaven’ featured in disciplines.

In this graph entities occurring in the analyzed novel are shown. They are extracted by the DBpedia Spotlight extractor. For every entity, a discipline (academic) is tried to be found within DBpedia articles and categories. The larger dots in various colours represent these disciplines. The graph is interactive.

More information is found on the project’s GitHub or in the abstract that was presented at the DHBenelux2016 conference.

More about this visualisation

Legend:
- DBpedia category
- Discipline

Search:

Search by name

http://dbpedia.org/resource/Category:History
Named Entity Recognition & Linking

- We are developing a new entity linker that allows for use of datasets other than DBpedia and is less sensitive to general entity popularity.

- Discovering more about Dark and NIL entities is also ongoing work.

Entity Typing using Distributional Semantics and DBpedia

Marieke van Erp and Piek Vossen
Vrije Universiteit Amsterdam
{marijke.van.erp,piek.vossen}@vu.nl
From words to concepts

• Linking terms to synonyms to obtain a higher level of abstraction

• Word-sense disambiguation + WordNet + Multilingual Central Repository + Framenet + PropBank

• Stop, quit, leave, relinquish, bow out -> all linked to the concept wn:leave_office
From Words to Concepts
Why link to WordNet/ConceptNet/etc?

- It allows you to query for types rather than instances: give me all lawsuits in the dataset

- In the context of CLARIAH, we are converting various diachronous lexicons to Linked Data
  - integrate resources
  - tag interesting concepts in text
  - query expansion
New synonym/concept lists are easy to plug in

Query expansion

Finding occupations in historic texts

‘s small farmers’

En van de schamelheid zijner plaggen had er de heikeuter nog eerst den langen weg te gaan tot de burgers van Venlo, eer hij de winst van zijn arbeid ingeruild zag tegen ’t noodige voor een schraal bestaan. (Felix Rutten, 1918, Ons mooie Limburg, DBNL)

Hisco
[occupation-65111-small farming]
kleinboer
kleinlandbouwer
keuterboer

Brouwers
[concept?]
keuterboer
heikeuter
landbouwer

 ..........
New synonym/concept lists are easy to plug in
Semantic Role Labelling

• Detecting the agent, patient, recipient and theme of a sentence

• Mary sold the book to John
  • Agent: Mary
  • Recipient: John
  • Theme: the book
Qatar Holding sells 10% stake in Porsche to founding families

Porsche family buys back 10pc stake from Qatar

fn:Commerce_money_transfer

type

fn:Buyer dbp:Porsche_family

fn:Goods

sem:hasTime 2013-06-17

fn:Money dbp:QatarHolding

Entity 10% stake

http://english.alarabiya.net
http://www.telegraph.co.uk
2013-06-17
Event abstractions

- Enable searches such as: Give me all lawsuits in which a politician was involved between 1990 and 2000.
Pragmatic Analysis

• Factuality/Attribution

• Who said what, who agrees with whom, how certain is a speaker about her statement, is she talking about the past, present or future?
Pro-EU campaigners have hoped that big carmakers would also support the Remain campaign.
and beyond...
Find out more

- All modules and evaluations are described in: http://kyoto.let.vu.nl/newsreader_deliverables/NWR-D4-2-3.pdf (158 pages!)
- http://www.newsreader-project.eu/results/software/
  - Black box setup
  - Links to individual modules on Github
  - Hadoop package for batch processing
Discussion

• It’s research software (no fancy interface)

• Currently not adapted to deal with old spelling variants/OCR/ etc

• NLP isn’t perfect (but humans don’t always agree either!)

• What would it take for you to start using such tools?

• What types of analyses are most interesting to the community?

• What use cases are most useful to the community at this point in time?
Thank you for your attention

https://youtu.be/rYLaVN3ogLI