Lightweight Text Analytics using Linked Data

Ali Khalili, Sören Auer, Axel-Cyrille Ngonga Ngomo

http://context.aksw.org
Agenda

- Motivation
- How does conTEXT work?
  - Workflow
  - Features
- Evaluation
- Conclusion
- Demo
Motivation: Analytical Information Imbalance
Motivation: Analytical Information Imbalance
Motivation: Analytical Information Imbalance

People should be able to find out what patterns can be discovered and what conclusions can be drawn from the information they share.
Motivation: Lightweight Text Analytics
Motivation: **Lightweight Text Analytics**

- IBM Content Analytics platform
- GATE
- Apache UIMA

Diagram:
- **Text Analysis Development Environment**
  - Text Analysis Tools
  - Business Intelligence Tools
  - Spreadsheets
  - Social Media Analysis Tools
  - Linked Data Analysis Tools

- **Flexibility of user interface**
  - Expert-programmer
  - Novice programmer
  - Non-programmer

- **Targeted user**

- **Unstructured**
- **Semi-structured**
- **Structured**
Motivation: Lightweight Text Analytics

- IBM Content Analytics platform
- GATE
- Apache UIMA
- Attensity
- Trendminer
- MashMaker
- Thomson Data Analyzer

Flexibility of user interface:
- Expert programmer
- Novice programmer
- Non-programmer

Unstructured
Semi-structured
Structured
Motivation: Lightweight Text Analytics

- IBM Content Analytics platform
- GATE
- Apache UIMA
- Attensity
- Trendminer
- MashMaker
- Thomson Data Analyzer
- Zoho Reports
- SAP NetWeaver
- Jackbe
- Rapidminer

Unstructured
Semi-structured
Structured
Motivation: Lightweight Text Analytics

- IBM Content Analytics platform
- GATE
- Apache UIMA
- Attensity
- Trendminer
- MashMaker
- Thomson Data Analyzer
- Zoho Reports
- SAP NetWeaver
- Jackbe
- Rapidminer
- Alchmey
- OpenCalais
- MashMaker
- Thomson
- Data Analyzer
- Zoho Reports
- SAP NetWeaver
- Jackbe
- Rapidminer
- Alchmey
- OpenCalais
- IBM Content Analytics platform
- GATE
- Apache UIMA
- Attensity
- Trendminer
- MashMaker
- Thomson Data Analyzer
- Zoho Reports
- SAP NetWeaver
- Jackbe
- Rapidminer
- Alchmey
- OpenCalais
Motivation: Lightweight Text Analytics

- IBM Content Analytics platform
- GATE
- Apache UIMA
- Attensity
- Trendminer
- MashMaker
- Thomson Data Analyzer
- Zoho Reports
- SAP NetWeaver
- Jackbe
- Rapidminer
- Alchmey
- OpenCalais
- Excel
- Data Wrangler
- Google Docs Spreadsheets
- Google Refine
- Alchmey
- OpenCalais
- Excel
- Data Wrangler
- Google Docs Spreadsheets
- Google Refine
Motivation: Lightweight Text Analytics

Text Analysis Tools
- IBM Content Analytics platform
- GATE
- Apache UIMA
- Attensity
- Trendminer
- MashMaker
- Thomson Data Analyzer
- Zoho Reports
- SAP NetWeaver
- Jackbe
- Rapidminer
- Alchmey
- OpenCalais

Linked Data Analysis Tools
- Facete
- CubeViz

Business Intelligence Tools
- Excel
- Data Wrangler
- Google Docs Spreadsheets
- Google Refine

Spreadsheets
- Google Docs Spreadsheets

Social Media Analysis Tools

NLP APIs

Text Analysis Development Environment

ConTEXT

Flexibility of user interface

Low
- Expert-programmer

Novice programmer
- Targeted user

Non-programmer

Unstructured

Semi-structured

Structured
Motivation: Lightweight Text Analytics

- IBM Content Analytics platform
- GATE
- Apache UIMA
- Attensity
- Trendminer
- MashMaker
- Thomson Data Analyzer
- Zoho Reports
- SAP NetWeaver
- Jackbe
- Rapidminer
- Alchmey
- OpenCalais
- Excel
- Data Wrangler
- Google Docs Spreadsheets
- Google Refine
- TweetDeck
- Topsy
- Flumes
Motivation: Lightweight Text Analytics

Text Analysis Tools
- IBM Content Analytics platform
- GATE
- Apache UIMA
- Attensity
- Trendminer
- MashMaker
- Thomson Data Analyzer

Text Analysis Development Environment

Linked Data Analysis Tools
- Facete
- CubeViz

Business Intelligence Tools
- Zoho Reports
- SAP NetWeaver
- Jackbe
- Rapidminer
- Alchmey
- OpenCalais

Spreadsheets
- Excel
- Data Wrangler
- Google Docs Spreadsheets
- Google Refine

Social Media Analysis Tools
- TweetDeck
- Topsy
- Flumes

NLP APIs

Flexibility of user interface
- Expert-programmer
- Novice programmer
- Non-programmer

Unstructured
- Semi-structured
- Structured
Motivation: Lightweight Text Analytics

Lack of tools dealing with unstructured content, catering non-expert users and providing extensible analytics interfaces.
conTEXT

- [http://context.aksw.org](http://context.aksw.org)
- A platform for lightweight text analytics

Approach

- No installation and configuration required
- Access content from a variety of sources
- Instantly show the results of analysis to users in a variety of visualizations
- Allow refinement of automatic annotations and take feedback into account
- Provide a generic architecture where different modules for content acquisition, natural language processing and visualization can be plugged together
A platform for lightweight text analytics

Approach
- No installation and configuration required
- Access content from a variety of sources
- Instantly show the results of analysis to users in a variety of visualizations
- Allow refinement of automatic annotations and take feedback into account
- Provide a generic architecture where different modules for content acquisition, natural language processing and visualization can be plugged together
Data Collection

- Handling different input types
Data Collection

- Handling different input types
Data Collection

- Handling different input types

Data transformation
- Rest APIs
- SPARQL endpoints
- RSS, Atom, RDF feeds
- Web Crawlers

Input Data Model
Data Collection

- Handling different input types

Input Data Model
- Relational
- RDF-based

Data transformation
- Rest APIs
- SPARQL endpoints
- RSS, Atom, RDF feeds
- Web Crawlers
Data Analysis
Data Analysis

- Natural Language Processing (NLP)
Data Analysis

- Natural Language Processing (NLP)

- DBpedia Spotlight
  [http://spotlight.dbpedia.org](http://spotlight.dbpedia.org)

- FOX
  [http://fox.aksw.org](http://fox.aksw.org)

- Any other NLP services which support NIF
NLP Interchange Format (NIF)

- [http://nlp2rdf.org](http://nlp2rdf.org)
- An RDF/OWL-based format
- Provides **Interoperability** between Natural Language Processing (NLP) tools and services.
- Standardize **access** parameters, **annotations** (e.g. tokenization), **validation** & **log** messages.
NLP Interchange Format (NIF)

“My favourite actress is Natalie Portman.”

- **Tokenizer**
  - `<#char=3,12>`
  - a nif:String, nif:RFC5147String, nif:Word;
  - nif:anchorOf “favourite”;
  - nif:referenceContext `<#char=0,>`;
  - nif:beginIndex “3”;
  - nif:endIndex “12”.

- **Snowball Stemmer**
  - `<#char=3,12>`
  - nif:stem “favourit”.

- **Stanford Core NLP**
  - `<#char=3,12>`
  - nif:oliaLink <http://purl.org/olia/penn.OWL#JJ>;
  - nif:oliaCategory <http://purl.org/olia/olia.owl#Adjective>;
  - nif:lemma “favorite”. [sic]

- **DBpedia Spotlight**
  - `<char=3,12>`
  - itsrdf:talentRef <http://dbpedia.org/resource/Favourite>;
  - itsrdf:taConfidence “0.10″^^xsd:decimal.

- **Integration through merged RDF**
  - `<#char=3,12>`
  - a nif:RFC5147String, nif:String;
  - a nif:Word;
  - nif:anchorOf “favourite”;
  - nif:referenceContext `<#char=0,>`;
  - nif:beginIndex “3”;
  - nif:endIndex “6”;
  - nif:stem “favourit”;
  - nif:oliaLink <http://purl.org/olia/penn.OWL#JJ>;
  - nif:oliaCategory <http://purl.org/olia/olia.owl#Adjective>;
  - nif:lemma “favorite”;
  - itsrdf:talentRef <http://dbpedia.org/resource/Favourite>;
  - itsrdf:taConfidence “0.10″^^xsd:decimal.

@base <http://example.org/prefix>
Data Enrichment
Data Enrichment

- **De-referencing the DBpedia URIs of the recognized entities.**

  (e.g. *longitude and latitudes for locations*, *birth and death dates for people*, etc.)
Data Enrichment

- De-referencing the DBpedia URIs of the recognized entities.
  (e.g. longitude and latitudes for locations, birth and death dates for people, etc.)

- Matching the entity co-occurrences with pre-defined natural language patterns for DBpedia predicates provided by BOA (BOotstrapping linked datA)
  (e.g. authorship relation)
Data Enrichment

- De-referencing the DBpedia URIs of the recognized entities.
  (e.g. longitude and latitudes for locations, birth and death dates for people, etc.)

- Matching the entity co-occurrences with pre-defined natural language patterns for DBpedia predicates provided by BOA (BOotstrapping linked datA)
  (e.g. authorship relation)
Data Mixing (Mashups)

- NLP service integration
- Composite corpus
  - E.g. Twitter + Blog + Facebook
- Helps to create a user model
Data Visualization & Exploration

- Different Views on Semantically-enriched data
  Using Exhibit & D3.js
### Faceted browsing

**Articles**
- Press Release: LOD2 Project Launch
- LOD at Energy Related Information Systems – an Interview.
- DBpedia Version 3.6 released.
- Linked Data for eGovernment – A Tutorial Session...
- DBpedia Knowledge Base Version 3.9

**Types**
- Place
  - PopulatedPlace
  - Country
  - Settlement
  - City
  - Town
  - (others)
- AdministrativeRegion
- Island
- Continent

**Entities**
- LOD2
- Open Data
- Linked Data
- DBpedia
- RDF
- LOD
- Semantic Web
- webinar
- open data
- Austria

**Articles Date**
- March 2012
- June 2011
- February 2012

**LOD2 project organises Linked (Enterprise) Data Track @ ESTC2010**
The 4th European Semantic Technology Conference 2010 (ESTC 2010) will take place in Vienna, Austria from 2nd to 3rd of December 2010. The STI International asked the LOD2 project consortium to organise a Linked (Enterprise) Data track / workshop on this years conference.

Although it is still very early in the LOD2 project lifetime (as we have kicked off the project in early September) the LOD2 team is happy to organise this track – we have bundled together an interesting series of presentations and talks for the ESTC2010 as follows:

**Linked (Enterprise) Data Track**
organised by LOD2 project (http://lod2.eu)
ESTC2010 – 02.12.2010, 10.30am to 12.30pm
- Linked Enterprise Data, business scenarios
  - Amar-Obai MEZAOUER, Exalex
- Big Data: experience of running the LOD warehouse of some 18bn triples
  - Orli Erling, OpenLink Software
- The role of SKOS in a Web of Data – some business use cases
  - Andreas Blumauer, Semantic Web Company (SWC)
- Licensing Issues of Linked (Open) Data (tbc)
  - Rufus Pollock, Open Knowledge Foundation

For detailed information on the track as well as on the above mentioned 4 presentations/talks please visit the respective ESTC programme page.

We are looking forward to meeting all interested people @ the ESTC2010 in Vienna – so save the date and register now!

Btw there is a Semantic Web meetup Vienna on the 1st December 2010 in the evening – so if you are in Vienna this evening – register for free & join us!

**DBpedia Version 3.6 released**
We are happy to announce the release of DBpedia 3.6. The new release is based on Wikipedia dumps dating from October/November 2010.

The new DBpedia dataset describes more than 3.5 million things, of which 1.67 million are classified in a consistent ontology including 364,000 persons, 462,000 places, 99,000 music albums, 54,000 films, 16,500 video games, 148,000 organizations, 148,000 species and 6,200 diseases. The DBpedia dataset features a total of 11.7 billion triples, and 3.5 billion links to external ontologies.
Places map & People timeline
Tag cloud
Chordal graph view
Matrix view
Annotation refinement

- Lightweight text analytics as an *incentive* for users to revise semantic annotations
- **RDFaCE** WYSIWYM (What-You-See-Is-What-You-Mean) interface for manual content annotation in RDFa format
- **Feedback** to NLP services → NLP calibration

DBPedia Spotlight Feedback API
http://spotlight.dbpedia.org/rest/feedback

FOX Feedback API
http://139.18.2.164:4444/api/ner/feedback
conTEXT architecture overview
Other features:
Interactive & Progressive Annotation

Interactive systems can be responsive despite low performance.
Other features:

**Real-time Semantic Analysis (ReSA)**

https://github.com/ali1k/resa
Other features:

• Search Engine Optimization (SEO) using Schema.org & JSON-LD

• Drilling down results using a subgraph of DBpedia

• Changing the underlying DBpedia ontology
Evaluation: Usefulness study

- Task-driven usefulness study
- 25 Users
- 10 questions pertaining to knowledge discovery in corpora of unstructured data
  - E.g. What are the five most mentioned countries by Bill Gates tweets?
Evaluation: Results of usefulness study

Measuring time & Jaccard similarity for answers using/wihtout conTEXT
Evaluation: Results of usefulness study

Avg. 136% more time without conTEXT

Measuring time & Jaccard similarity for answers using/without conTEXT
Evaluation: **Usability study**

- **System Usability Scale (SUS)**
  - Score: 82

http://www.measuringusability.com/
Evaluation: Usability study

- System Usability Scale (SUS) → 82

http://www.measuringusability.com/
Conclusions

Lightweight Text Analytics using Linked Data

- Democratizing the NLP usage
- Alleviating the Semantic Web's chicken-and-egg problem
- Harnessing the power of feedback loops
Future Work

- Improving the **performance & scalability** of views
- Exposing **APIs** for third-parties
- Enable **batch** refinement of annotations
- More **input** source types
- More…
Any Questions?
Demo

- Progressive data collection and annotation
  - [http://context.aksw.org](http://context.aksw.org)
- Different views
  - LOD2 Blog
- Example of adding extra input types + changing the DBpedia ontology + composite corpora
  - LinkedIn Jobs