Can we Produce Multilingual NLP Workbenches for DH Researchers? Report on the LitText Experiment

Andrew U. Frank

Geoinformation
TU Vienna

frank@geoinfo.tuwien.ac.at

Outline

Message:

Building NLP workbenches for DH researchers should be "multilingual".

- NLP tools are ready for users.
- More information about NLP tools for multilingual setup is required.

My background

- Dipl.Ing.(surveying and mapping) and
- Ph.D.(databases) from ETH Zurich.
- Professor for Geographic Information Science at
  - U Maine and
  - TU Wien.
- Member of the founding team of the NSF funded National Center for Geographic Information and Analysis.
- Research focus on spatial information and cognition, cognitive linguistics.

Experiment to explore limitations

Concrete experiments to build systems for DH help to assess

- the current state of the art and
- the limitations for use of NLP tools

by DH researchers.

Workbench for Computation Comparative Literature studies

Traditional (manual) literary studies proceed in steps:

- collect the texts in several languages to analyze
- read and annotate the texts
- build the corpus
- extract the relevant parts.

The same approach should be carried over to a digital workbench.

Example Analysis

As an example for a computational comparative literature analysis:

Find all sentences in novels where animals behave like humans, i.e.

- think rationally,
- communicate specific with language etc.

Source Texts are files

- The source texts are each a file and organized in folders.
- Descriptions (metadata) are stored in the source files. This reduces complexity,
  - no differences between source and descriptions
  - no additional support for managing metadata required.
- Simple markup language for
  - metadata,
  - layout of text on pages and
  - text structure in chapters and paragraphs.
- Open to connect with other useful data collections.

Test with texts from Project Gutenberg

Moretti 2007 with his now famous book "Graphs, Maps, Trees: Abstract Models for Literary History"
started a trend for DH to study novels of the 19th century:

- Texts are available and free of copyright.
- Large selection, mostly English, but other languages as well.

A large number of novels available

- >10,000 English,
- 500 - 1,500 for German, French, Italian, Spanish.

NLP tools as service

We started with the Stanford coreNLP pipeline:

- easy to setup for English
- other languages available.

NLP services are setup as servers at URL using the http protocol.

Manage corpus as Linked Data triple store

RDF triples are

- flexible,
- extensible and
- schema-less

Triple stores to manage data collection.
Sound theoretical base: (binary) relation, category theory.

Multiple formats with fast and loss-less transformation (typically compacted wtih zip).

Analyze each text individually and store results as triples (as nt.gz files).

Triple Data Structure

Linked Data is structured as triples subject - property - object, where

- subject is an identifier,
- property is a description of the relation between subject and object,
- object is a value or another identifier.

  Example
  <http://globalwordnet.org/wordnets-in-the-world/> rdf:about "wordnet" .

Identifiers and property names are globaly unique IRI (aka URL).

SPARQL query language

SPARQL in version 1.1 is a standardized language to manage and query RDF triple data collections.
Several open-source and commercial triple stores to load triples and query with SPARQL (we use Jena with Fuseki).
SPARQL is (essentially) a simplified SQL (because all relations are binary).

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX nlp: <http://gerastree.at/nlp_2015#>
select *
from <http://gerastree.at/test251m>
where {
?tok nlp:wordForm "bachelor"@eng .
?tok rdfs:partOf ?sent .
?sent nlp:sentenceForm ?sentForm .
}
```

Character encoding schemes

Everybody is using UTF-8!
but there are

legacy data e.g.
Project Gutenberg delivers UTF-8 encoded files, occasionally the conversion from older formats went wrong.
Example an Italian novel contains
"nč solo perde ciň che possedeva di bello e di grande, ma cade nel piů profondo della miseria e del languore;"

legacy researchers use (non-current) Windows or Apple Macintosh OS and ignore character encodings; transformations are too often transparent to the user...

Language of the markup language

Keywords for the markup language:
English enforced or in the language of the researcher?
Title or Titel, PublicationDate or ErscheinungsDatum?
Language, Sprache, ...

Which languages are supported

Stanford's web page lists the language for which trained models are available

LANGUAGE	MODEL JAR	VERSION
Arabic	download	3.8.0
Chinese	download	3.8.0
English	download	3.8.0
English (KBP)	download	3.8.0
French	download	3.8.0
German	download	3.8.0
Spanish	download	3.8.0

It does not say, what specific annotators are available for each language.
For example, German does not produce lemmata but coreferences,
other languages seem not to have coreference annotators.
Input encoding seem to be now standardized for UTF-8
- progress is made!.

Translation to Triples:

<http://gerastree.at#test-t10/P00001/N01/S000001>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://gerastree.at/nlp_2015#Sentence> ;
<http://gerastree.at/nlp_2015#parse> "(ROOT\n (NP\n (NP (DT The) (NNP Decameron))\n (PP (IN of)\n (NP (NNP Giovanni) (NNP Boccaccio)))\n (. .)))\n\n" ;
<http://www.w3.org/2000/01/rdf-schema#partOf> <http://gerastree.at#test-t10> ;

<http://gerastree.at/nlp_2015#sentenceForm> "The Decameron ."@eng .
<http://gerastree.at#test-t10/P00001/N01/S000001/T001><http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://gerastree.at/nlp_2015#Token> ;
<http://www.w3.org/2000/01/rdf-schema#partOf>
<http://gerastree.at#test-t10/P00001/N01/S000001> ;
<http://gerastree.at/nlp_2015#lemma3> "the"@eng ;
<http://gerastree.at/nlp_2015#pos> "DT" ;
<http://gerastree.at/nlp_2015#wordForm> "The"@eng ;

<http://gerastree.at/nlp_2015#speakerTag> "PER0" .
<http://gerastree.at#test-t10/P00001/N01/S000001/T002>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://gerastree.at/nlp_2015#Token> ;
<http://www.w3.org/2000/01/rdf-schema#partOf>
<http://gerastree.at#test-t10/P00001/N01/S000001> ;
<http://gerastree.at/nlp_2015#lemma3> "Decameron"@eng ;
<http://gerastree.at/nlp_2015#pos> "NNP" ;
<http://gerastree.at/nlp_2015#wordForm> "Decameron"@eng ;
<http://gerastree.at/nlp_2015#speakerTag> "PER0" .

POS and similar codes

Conversion to triples requires a "vocabulary".
Currently strong adherence to coreNLP format, using the field names as property names.
The structure of output seems uniform across languages, as long we use we use coreNLP based tools.
No experience yet with other NLP tools.

Difficulties with codes: POS are (currently) language specific and it is difficult to find the authoritative descriptions.
Tools to discover the codes actually used.
UD is a step in the right direction, but again missing documentation:
What is e.g. "NMOD+Missing "TOWARD"" meaning?

Additional Services

Lemmatization is crucial for most DH research but missing with the German model in coreNLP.
An additional service for lemmatization of German input is constructed from Schmid's TreeTagger.

Package of tools as services

coreNLP is packaged to run as a server.
Only a setup for automatic starting and assignment of a port for each language service is needed.
Linux systemd can be setup to start and restart the service.
Other tools require a wrapper:
For example, Schmid provided me a TreeTagger as a service.
It uses a socket and the wrapper needs only translate from the http port to the socket.

SPARQL is likely to complex for DH researchers

The base structure of SPARQL is simple and easy to learn.
A query is a path through the web of connections between the triples.
The complexity comes from the "vocabularies" (i.e. the names for the properties).
In addition, the differences between the encodings for different languages (POS, NER, Speakers, Dependencies) make queries crossing language boundaries very challenging.

List of subordinates (hyponyms)

For the example query extracting animals which think, requires

- a list of all subordinates of "communicate" and "think" (i.e. wordnet's "cerebrate")
- as well as a list of all subordinates of "animal" are required.

Wordnet is available from Princeton in a Linked Data format. It can be loaded in a triple store and connected with the literary data.

Obtaining wordnet like datasets in other languages varies. A useful list is maintained by http://globalwordnet.org/wordnets-in-the-world/. For some languages the data can only be obtained after a lengthy licensing process (often free of charge for academic use).

Download and markup

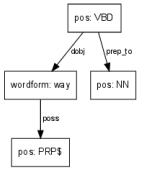
Download and automatic markup is very fast (few seconds per novel);
automatic markup is not precise enough, due to slight differences in the format of the files of Project Gutenberg (mostly a legacy issue).
Manual editing and completing markup takes few minutes per title).
On average, a novel takes less than 1 MB storage.

NLP analysis

On dated hardware
(i5 with 4 cores at 3 Ghz, running Linux and openJDK 1.8)
the processing of novels with 1.2 million words takes 40 minutes.
The resulting nt.gz files are relatively small
(10 times the original text file, going from 7.2 to 70 MB)

Queries

Queries closely following the queries used in a benchmark paper by Proisl and Uhrig in 2012 gives extremely (< 1sec) responses
but strongly dependent on caching the data.
- with a corpus 1/5 of the corpus used originally (momentary lack of disk space to build a larger one).




```
select *
from <http://gerastree.at/test251m>
where {
?tok nlp:wordForm "way"@eng .
?dep nlp:governor ?tok ;
          nlp:dependency "NMOD+POSS" ;
          nlp:dependent ?tok2 .
?tok2 nlp:pos "PRPdollar" ;
          nlp:wordForm ?wf2 .
?dep2 nlp:dependent ?tok ;
          nlp:dependency "DOBJ" ;
          nlp:governor ?tok3 .
?tok3 nlp:lemma3 ?lem4 ;
          nlp:pos "VBD" .
          ?dep3 nlp:governor ?tok3 ;
          nlp:dependency "NMOD+TO" ;
          nlp:dependent ?tok4.
?tok4 nlp:pos "NN" ;
          nlp:lemma3 ?lem3 .
?tok rdfs:partOf ?sent .
?sent nlp:sentenceForm ?sf .
}
```

Building Multilanguage Workbenches for DH is possible

- Tools are available,
- hardware is fast and
- inexpensive enough,
- performance is acceptable.

Better documentation of what a tool delivers, including TreeTag codes and available annotators.

Wordnet like datasets for many languages.

The information is likely available, even on the web, but only for language specific communities, who know the right keywords.

- How difficult is it to produce a model for coreNLP?
- There must be training corpora available for many languages, used to produce the tools listed (e.g. TinT for italian)?
- What could be done to advance multilanguage tools? A