Sentiment Analysis and Visualization using UIMA and Solr

Carlos Rodríguez Penagos, David García Narbona, Guillem Massó Sanabre, Jens Grivolla, Joan Codina Filbà
Sentiment Analysis

- Social Media Monitoring, Reputation Management, Opinion Mining, ...

- “Who says what about what?”

- or “What do people say about my product/brand?”
Product Review Analysis

- **Objective:** analysing customer opinion from unstructured product reviews

- **Approach:**
  - detect Opinionated Units (Targets and Cues) → UIMA
  - data mining / visualization of target-cue relations → Solr, Cluto, etc.
Architecture Overview
Architecture Overview (detail)

Review & Database

Linguistic annotation
- POS, lemmas, NER, Chunks, dependencies, etc.

Target & Cue detection

Opinionated Unit detection
- T&C Correlation via dependencies

Data visualization

OU indexing

OU polarity Assignment
OU detection

- combine statistical and rule-based approaches
  - reliably find known entities and opinion expressions
  - discover new entities and opinions
OU detection

- mark known Targets (e.g. brand / product names, etc.) and known Cues (e.g. polar words and expressions)
- detect new Targets and Cues using statistical models
- relate Targets and Cues through syntactic dependencies
OU detection
Visualization

- using Ajax-Solr
- all data preprocessed and indexed with Solr
- flexible interactive querying/filtering
- clustering using Carrot, Cluto, Solr-based kNN, etc.
http://webmining.barcelonamedia.org/sm_yahoo/
Some results

- participated in SemEval (assigning polarity to tweets) with good results:
  - 5th out of 23 submissions
  - 0.86 avg. F1 measure
- customer review corpus (manually annotated at BM):
  - 88.5% correctly identified OUs
  - 70% correct polarity
UIMA: challenges

- combining components from different sources (and languages: Java, C++, Python)
- unified Type System
- non-programmers need to create pipelines and AEs
UIMA components

- OpenNLP (Apache)
- JNET (JulieLabs)
- Zanzibar (Tor Vergata University)
- Lemmatizer (BM)
- DeSR (University of Pisa, wrapper by BM)
- DependencyTreeWalker (BM)
- Weka Wrapper (based on MAWUI by Mayo Clinic)
- UIMA Collection Tools (BM)
OpenNLP

- no code changes
- already TS independent
- just add XML descriptor + resource (model)
JNET

- major code changes
- made TS independent
- fixed bugs related to rich feature vectors
- would be nice to merge upstream
Zanzibar

- used for NP detection
- major code changes / bug fixes
- upstream?
  - seems mostly abandoned (2011)
- probably move over to RUTA
Lemmatizer

- uses ConceptMapper to generate all possible lemmas
- custom module to filter candidates by POS tag
DeSR

- wrapper for the DeSR parser ([https://sites.google.com/site/desrparser/](https://sites.google.com/site/desrparser/))
- developed using UIMA-CPP
- developed at BM
- available on GitHub ([https://github.com/BarcelonaMedia-ViL/desr-uima](https://github.com/BarcelonaMedia-ViL/desr-uima))
DependencyTreeWalker

- developed at BM
- uses Pythonnator
- enables lookups in the dependency graph
- used to validate Target-Cue relations
Weka Wrapper

- based on MAWUI
- many changes
  - adapted to newer UIMA versions
  - bug fixes, ...
- upstream not updated since 2008
- our own changes not published so far
Configurable Annotator

- taken from LuCAS (Apache UIMA)
- preprocessing / extraction as a separate module (without lucene dependency)
- used to prepare annotations for WEKA and Solr
UIMA Collection Tools

- mostly based on example CRs and CCs from UIMA
- use MySQL (or Solr) instead of files
  - CR: plain text and XMI
  - CC: flat DB row representation or XMI
  - annotation viewer: works with XMI from DB
- developed at BM
- published on GitHub
What we do well

- separation of code and configuration
- type system independence of code
- managing code and components with git and maven
What we need to do/learn

- better resource handling (maven?)
- avoid redundancies between code and descriptors (uimaFIT?)
- automatize creation of new components (e.g. variants using other models)
- publish our changes
  - github
  - upstream
- integrate a better rule engine (Ruta?)
- better separation of libraries, etc. for CPP or Python annotators
New EU (FP7) project: EUMSSI

“Event Understanding through Multimodal Social Stream Interpretation”

⇒ using UIMA as an integration platform for multimodal analysis layers

starts December 2013