Partners

- **The University of Sheffield**
  
  GATE team, Department of Computer Science
  
  Regent Court, 211 Portobello, Sheffield S1 4DP, UK
  
  PI [Dr Wim Peters](mailto:wim.peters@sheffield.ac.uk) | Co-PI [Dr Kalina Bontcheva](mailto:kalina.bontcheva@sheffield.ac.uk) | Co-PI [Dr Mark Hepple](mailto:mark.hepple@sheffield.ac.uk)

- **MODUL University Vienna**
  
  Department of New Media Technology
  
  Am Kahlenberg 1, 1190 Vienna, Austria
  
  PI [Prof Arno Scharl](mailto:arno.scharl@univie.ac.at) | Co-PI [Dr Marta Sabou](mailto:marta.sabou@univie.ac.at)

- **Vienna University of Economics and Business**
  
  Research Institute for Computational Methods
  
  Augasse 2-6, 1090 Vienna, Austria
  
  PI [Prof Kurt Hornik](mailto:kurt.hornik@univie.ac.at) | Co-PI [Dr Gerhard Wohlgenannt](mailto:gerhard.wohlgenannt@univie.ac.at)

- **LIMSI-CNRS**
  
  Man-Machine Communication Dept, ILES group (Written and Signed Language)
  
  Bat. 508, Univ. Paris-Sud, 91403 Orsay Cedex, France
  
  PI Dr [Patrick Paroubek](mailto:paroubek@limsi.fr) | Co-PI Dr [Amel Fraisse](mailto:amel.fraisse@limsi.fr)
uComp Objectives

- Develop a generic, configurable and reusable human computation framework
- Address challenges of noisy data
- Embed human computation into knowledge extraction workflows
  - Factual Knowledge
  - Affective Knowledge
- Evaluate EHC performance
  (EHC = Embedded Human Computation)
Data Acquisition

- **Unstructured Data**
  - Start: 01 Jan 2013
  - Filter: Climate Change
  - Social Media Postings: 10,333
  - News Media Articles: 24,605
  - Extensible Web Retrieval Toolkit (eWRT)
    Open Source Library | www.ecoresearch.net/ewrt

- **Structured Data**
  - Evaluation of Triple Stores
  - uComp Linked Data Repository (Sesame)
  - Data dumps of generic resources: DBpedia, GeoNames
Extensible Web Retrieval Toolkit (eWRT)

Knowledge capture in the age of massive Web data requires robust and scalable mechanisms to acquire, consolidate and pre-process large amounts of heterogeneous data. The Extensible Web Retrieval Toolkit (eWRT) is modular open-source Python API that addresses this requirement. It retrieves social data from Web sources such as Delicious, Flickr, Yahoo! and Wikipedia, including various helper classes for effective caching and data management. The toolkit provides components for content acquisition and caching, low-level natural language processing functionalities such as language detection, phonetic string similarity measures, and methods for string normalization.

eWRT has been jointly developed by researchers from MODUL University Vienna, weblyzard technology, the University of Applied Sciences Chur, and the Vienna University of Economics and Business. The library is currently being extended as part of the uComp Project, which investigates Embedded Human Computation for Knowledge Extraction and Evaluation.
Mining Noisy Textual Data

- **Social media content**, and especially tweets, is problematic to mine, because sentences are very short and/or incomplete.
- **Linguistic tools** such as Part-of-Speech (POS) Taggers and Named Entity Recognition (NER) do not perform well:
  - Stanford POS tagging accuracy drops from about 0.97 for news to 0.80 for tweets (Ritter, 2011)
  - Stanford NER 62% (Ritter, 2011)
- Even usually straightforward methods like **language identification** can have problems.
TwitIE: IE on Social Media

• Adapting the **ANNIE** open-source NLP system from GATE to processing noisy content
  • Normalisation
  • Tweet-tailored Tokenisation (@user, #tag)
  • POS tagging
  • Named Entity Recognition
• First results published in a conference paper, to appear at Hypertext 2013
• Open-source **TwitIE Release** forthcoming in May 2013
TwitIE: IE on Social Media

Input: URL or text

Document format (JSON, XML, HTML, etc)

GATE Document

Language Identification

Social-media specific language footprints

TwitIE Tokeniser

Character Class Sequence Rules

Gazetteer Lookup

Lists of Names (e.g. days of week)

Sentence Splitter

JAPE Sentence Patterns

Normaliser

Spelling and orthographic dictionary

Stanford POS Tagger

Twitter Adapted Model

Named Entity Recogniser

JAPE Grammar Cascade

Output: GATE Document XML dump of IE Annotations
• **Crowdsourcing application**, building upon a social application framework to engage users while generating valuable information.

• Players score if *inputs match*: (i) system-generated values; (ii) Real-time input from other players; (iii) stored records from previous users.

• If a certain number of players *agree*, the task will be assumed *complete* and taken out of the game.

• Tasks: **HTML5 application** to ensure compatibility with mobile platforms; provision of an **Application Programming Interface (API)**; experiments to integrate games with **CrowdFlower data**.
Sentiment Quiz

Is the following a negative, neutral or positive statement about the candidate?

"We are headed down a path that is certain to end in the destruction of our experiment in democracy."

Sentiment Quiz

Hat der folgende Ausdruck eine negative, neutrale oder positive Bedeutung?

amoralisch

Adjektiv

Status

Election Member

Tell your Friends!

Spread the Word

You will earn 10% of your friends' points after they accept your invitation. If you like the Sentiment Quiz, please become a fan and check out our Fanpage www.chistera.eu.
Affective Knowledge

- Use HC to produce affective resources that are difficult to obtain automatically and too costly to produce manually, for multiple languages (EN, FR, DE).
- Assess HC-produced resources by evaluating the performance impact of using them instead of traditional resources for opinion mining and sentiment analysis (quantitative black-box methodology).
- Assess the possibility to replace static gold standard resources by dynamic HC.
Factual Knowledge

- Ontologies create *shared meaning* and are a cornerstone of the *Semantic Web*
- *Manual construction* of ontologies is cumbersome and expensive
- *Ontology learning* is a (semi-)automatic process to assist the ontology engineer
- uComp builds on an existing ontology learning *framework*
Ontology Learning & HC

uComp aims to...

- support various subtasks of OL
- evaluate results from automatic processes on the concept, relation and instance level
- embed HC into the algorithms, adapting them based on the HC-provided feedback
- build a generic HC platform to facilitate the integration of additional steps in the ontology learning and verification cycle
- use multiple evidence sources (requires to evaluate their quality and assign source impact values)
Dissemination Results

• Project Web Site
  www.ucomp.eu

• Twitter Presence
  @uCompEU

• uComp is co-organising a Workshop on Scalability in NLP, collocated with the International Conference on Recent Advances in NLP (https://sites.google.com/site/scanlp2013)

• Tutorial Proposal: Mining Noisy Textual Data
  Annual Meeting of the Association for Computational Linguistics (ACL-2013)

• New Module on Mining Social Media with TwitIE
  Annual GATE Summer School (03-07 June 2013)
Publications

• Accepted

• Under Review
  • Knowledge Capture from Multiple Online Sources with the Extensible Web Retrieval Toolkit (eWRT), K-CAP 2013
  • Games with a Purpose or Mechanised Labour? A Comparison of Two Human Computation Genres. Sabou, M, Bontcheva, K., Scharl, A. and Foels, M. Submitted to SoHuman 2013. 2nd Workshop on Social Media for Crowdsourcing and Human Computation.
  • Improving Minor Opinion Polarity Classification with Named Entity Analysis, A. Fraisse & P. Paroubek, Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, 06/2013.
  • L’apport des Entités Nommées pour la Classification des Opinions Minoritaires, A. Fraisse & P. Paroubek, TALN, 06/2013.

• In Preparation
  • TwitIE Demo Paper, ACL 2013 (to be submitted 08 April 2013)