brat: a Web-based Tool for NLP-Assisted Text Annotation

Pontus Stenetorp[1]*, Sampo Pyysalo[2,3]*, Goran Topić[1], Tomoko Ohta[1,2,3], Sophia Ananiadou[2,3] and Jun'ichi Tsujii[4]

[1]Department of Computer Science, University of Tokyo | {[2]School of Computer Science, [3]National Centre for Text Mining}, University of Manchester | [4]Microsoft Research Asia

{pontus,smp,goran,okap}@is.s.u-tokyo.ac.jp, sophia.ananiadou@manchester.ac.uk, jtsujii@microsoft.com

* These authors contributed equally to this work

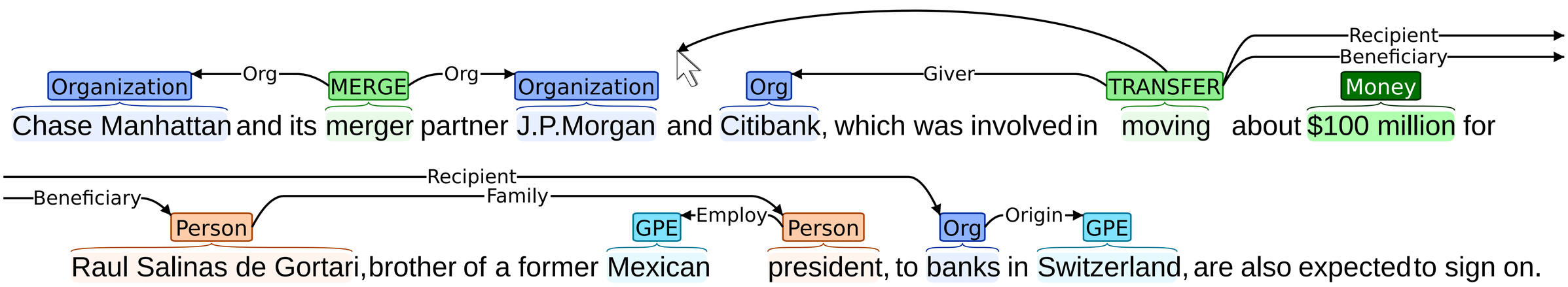

Figure: Create annotations using intuitive mouse gestures such as dragging and dropping.

brat rapid annotation tool (brat)

- Web-based annotation tool
- Intuitive corpus annotation
- Fully configurable
  - Text spans, relations, attributes, ***n***-ary associations, etc.
  - Support tasks ranging from POS tagging to event annotation
- Actively used and developed by several groups
- Fully **open-source** (MIT) and **available now**

Advanced Features

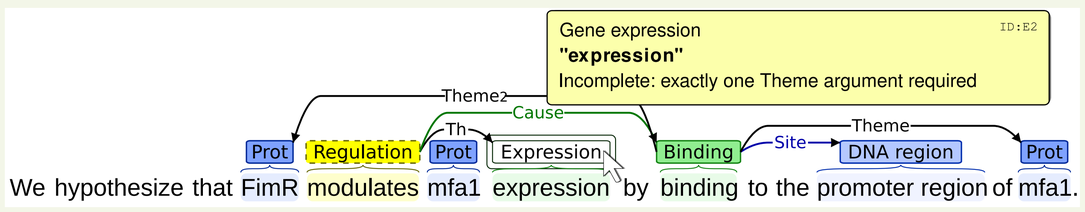

Figure: Problematic/incomplete cases marked with visual cues, details in pop-up.

- Annotation verification shows semantic constraint violations
- Fully featured search with detailed constraints
- Key-word-in-context concordancing
- Integration with automatic taggers as web services

Rapid Annotation Mode

- Multi-category annotation
- *Goal:* Limit number of categories in the UI exposed to the annotator
- *15%*/*31%* decrease in annotation/type selection

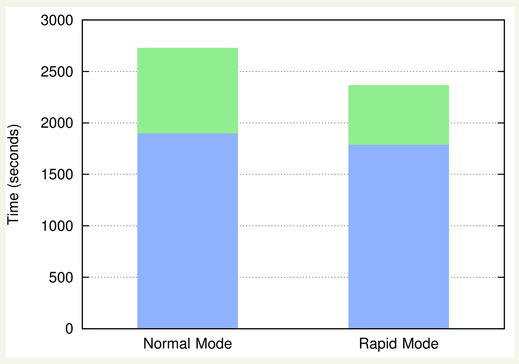

Figure: Annotation time. Green is time spent selecting annotation type.

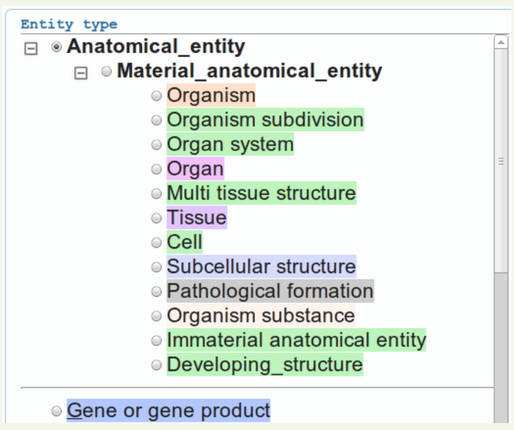

Figure: Normal mode.

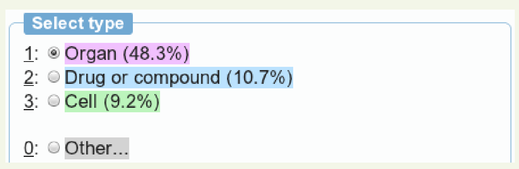

Figure: Rapid mode.

- Show most probable types
- Reduced UI cluttering
- Reduced hand/eye movement
- Surprisingly efficient

Architecture

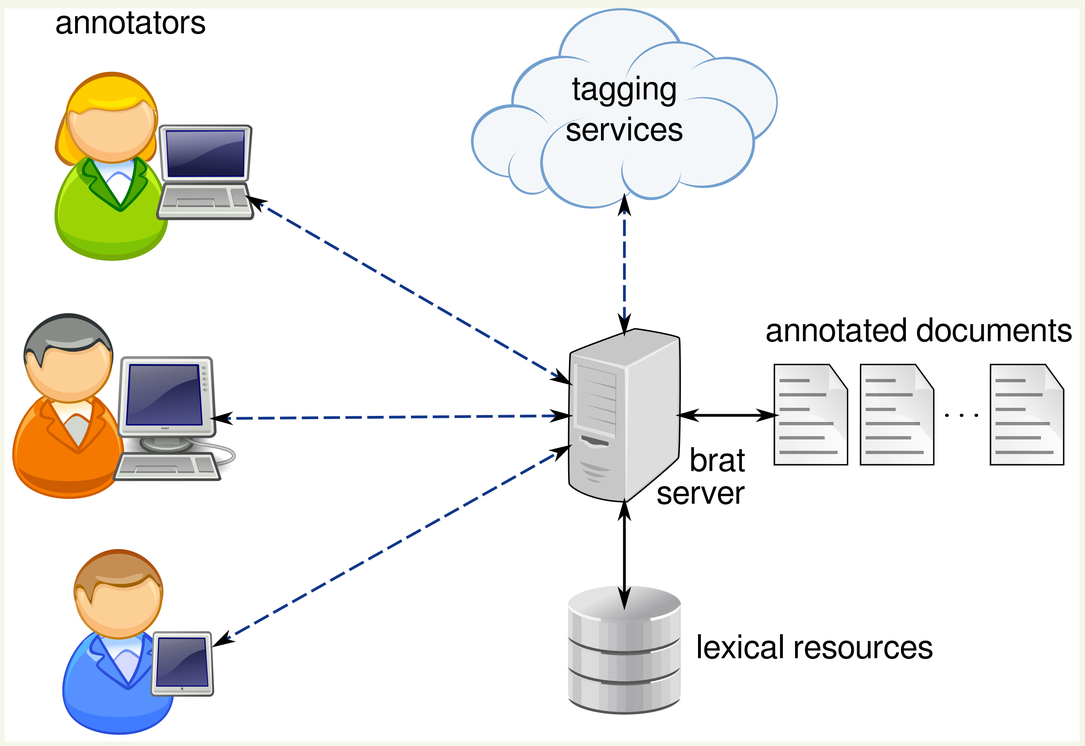

Figure: Client-server architecture with web-browser clients

- Server stores annotations and manages annotator edits
  - Always up-to-date, no chance of conflicting versions
  - No separate "save" operation, no data loss
  - Real-time collaboration possible
- "Put your eggs in one basket and watch that basket!" – Mark Twain

Annotation Format Details

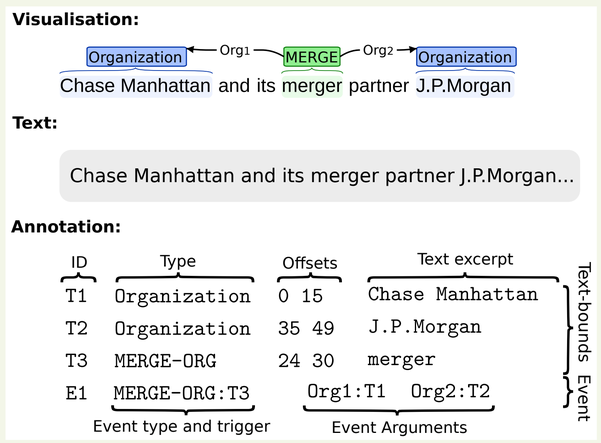

Figure: Stand-off annotation format

Annotation primitives:

1. Typed spans
2. Binary relations
3. ***N***-ary associations
4. Multi-valued attributes
5. Free-form text "notes"

- Stand-off format, allowing over-lapping annotations
- Conversion scripts for several popular formats included, such as BIO-tagged text, CoNLL-X/Malt dependencies, etc.

Availability

Live demonstration, source code, documentation, installation script, conversion tools, future additions and more is/will be available at:

http://brat.nlplab.org/

The 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012), Avignon, France, April 26th 2012