BLAH5: improved dependency annotation webservice using NER

- Proposal: dependency parser service improved by NER
- Side Discussion: Comparison of bio-NER tools / services

NER & dependency parsing

- OGER++ is a NER tool developed by UZH
- Simple interface, also useable as a web service
- Dictionary-based NER, using an aggregator service (BioTermHub)

Comparison

- Compared OGER, TaggerOne and Jensen Tagger (and informally also MetaMap)
- Comparison
  - MEDIC termlist provided by CTD (about 10 000 diseases and 68 000 synonyms) (for annotation quality)
  - 30 000 articles from Medlease (for speed)

Comparison: Termlist Selection

- OGER and Jensen Tagger take any termlist
- Tagger One needs to be trained on a given termlist.
- MetaMap has several limitations:
  - Requires UMLS authentication, terminology defined as an UMLS subset, accepts only ASCII characters

- To facilitate comparison, we used the term list that Tagger One is trained on by default.

- However, if you want to use it with your own term list, you need to retrain it.

TaggerOne

- Can be run without installation : download, read instructions, run

Bioinformatics. 2016 Sep 15;32(18):2839-46. doi: 10.1093/bioinformatics/btw343. Epub 2016 Jun 9.

TaggerOne: joint named entity recognition and normalization with semi-Markov I

Leaman R[1], Lu Z[1].

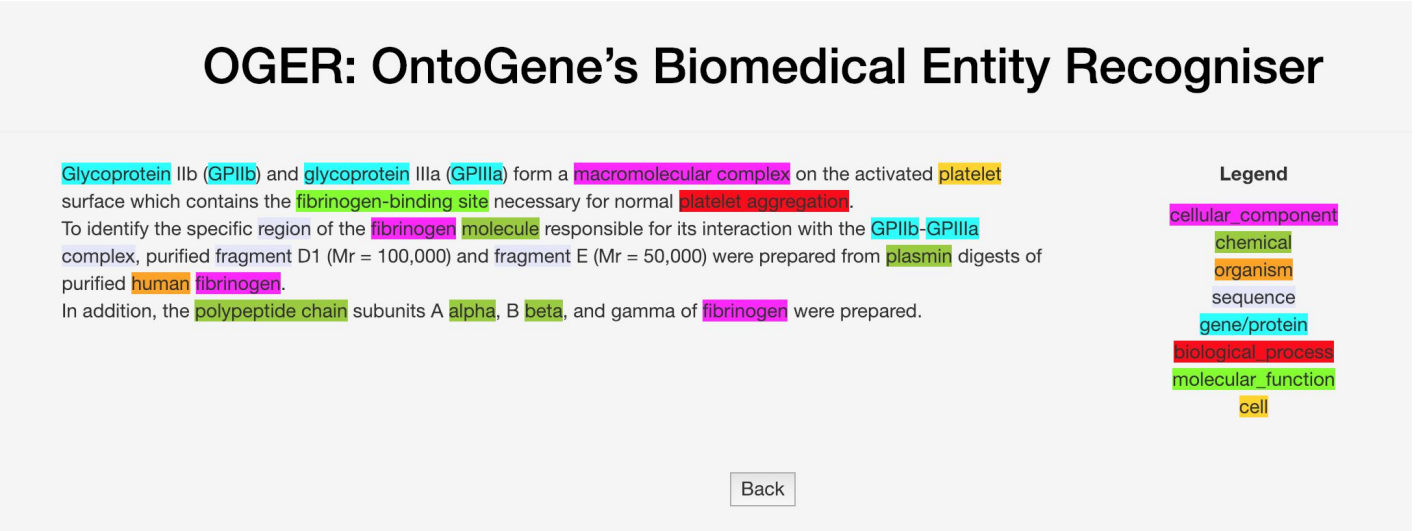
Author information

Abstract

MOTIVATION: Text mining is increasingly used to manage the accelerating pace of the biomedical literature. Many te
depend on accurate named entity recognition (NER) and normalization (grounding). While high performing machine
trainable for many entity types exist for NER, normalization methods are usually specialized to a single entity type. N
systems are also typically used in a serial pipeline, causing cascading errors and limiting the ability of the NER syste
lexical information provided by the normalization.

OGER++ (www.ontogene.org/)

- Can be run without installation : download, read instructions, run



Jensen Tagger

- Moderately easy to install, needs swig and boost libraries
- Termlists and input need to be in non-standard TSV format, so some conversion is needed
- Written in C++

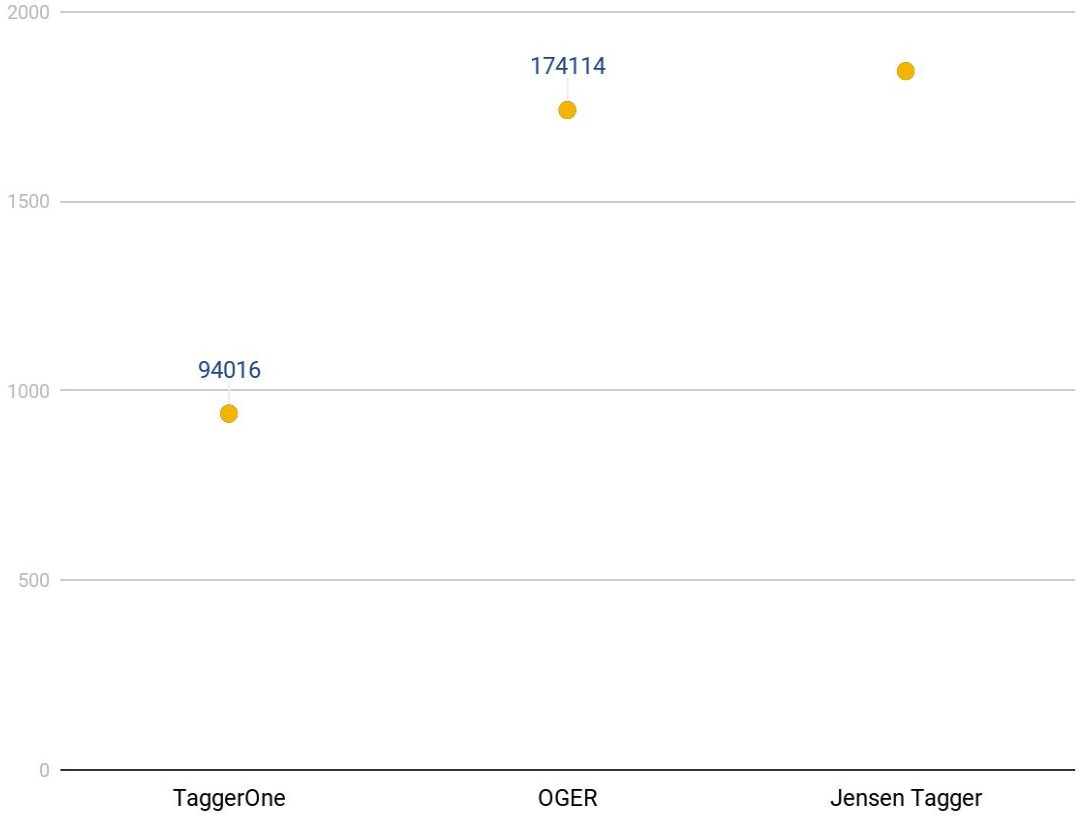
```
53
54  DocumentTagger::~DocumentTagger()
55  {
56  }
57
58  void DocumentTagger::load_names(const char* entities_filename, con
59          Tagger::load_names(entities_filename, names_filename);
60          this->entity_type_map = new EntityTypeMap(entities_filenam
61  }
62
63  void DocumentTagger::load_names(int type, const char* names_filena
64          Tagger::load_names(type, names_filename);
65  }
```

Comparison: Speed

- Note: OGER depends on NLTK components. Switching off stemming increases the speed and makes it faster than J. tagger
- MetaMap would take about one day, according to the author's declared speed: 22 abs/minute.

Comparison: Annotations

- The number of annotations is not necessarily an indicator for their quality
- If stemming in OGER is deactivated, the number of annotations decreases but not substantially



Comparison

Intrafamilial transmission of pulmonary tuberculosis due to Mycobacterium bovis.

- All three taggers (TaggerOne, OGER, Jensen Tagger) deal well with simple cases

Comparison

... causes a marked reduction in immune response to the novel antigen phytohaemagglutinin (PHA) …

Although individual anoles varied considerably in the extent to which they responded to PHA challenge,

our results suggest that an immune response can impose a substantial metabolic cost

- All three taggers also, being termlist-based, make the same mistakes (PHA here is not a disease)
- Jensen Tagger doesn't filter easy FPs natively (but can be added with a blacklist)

Comparison

4-Oxo-2-nonenal (ONE) … generated in increased amounts during degenerative diseases and cancer.

We show that pyridoxamine ... form pyrrolo[2,1-b][1,3]oxazines with the participation of both the amino and the phenolic groups.

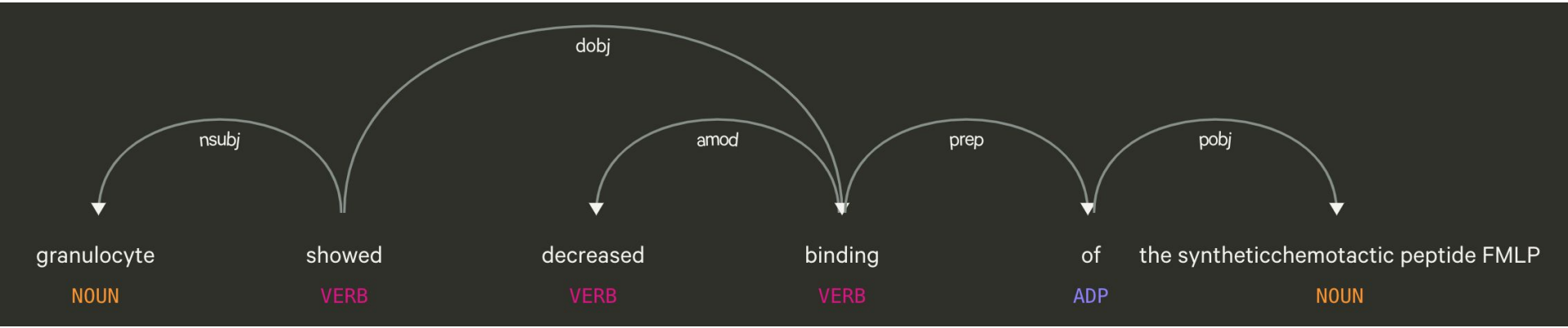
- OGER gets overambitious in chemical formulas, mapping 1 to *febrile seizures*, *familial*, *1* here

Proposal for BLAH5

- Develop a web service using spaCy accessible through a REST-API
  - spaCy is a state of the art NLP library, offering PoS tagging and dependency parsing
- The results of dependency parsing can be improved by *first* providing NER using OGER or any other service
- Service allows uploading some pre-annotated text, and receive improved annotations

Proposal for BLAH5

- providing dependency parser with information about what is a NER improves parse



Proposal for BLAH5

- Service allows uploading some pre-annotated text, and receive improved annotations
  - If no annotations are provided, service will ask NER service to provide annotations which to use to create improved dependency annotations.
- Independent, can be used for further processing such as relation extraction
- Ideally in JSON format compatible to PubAnnotation in the spirit of interoperability

Proposal for BLAH5