

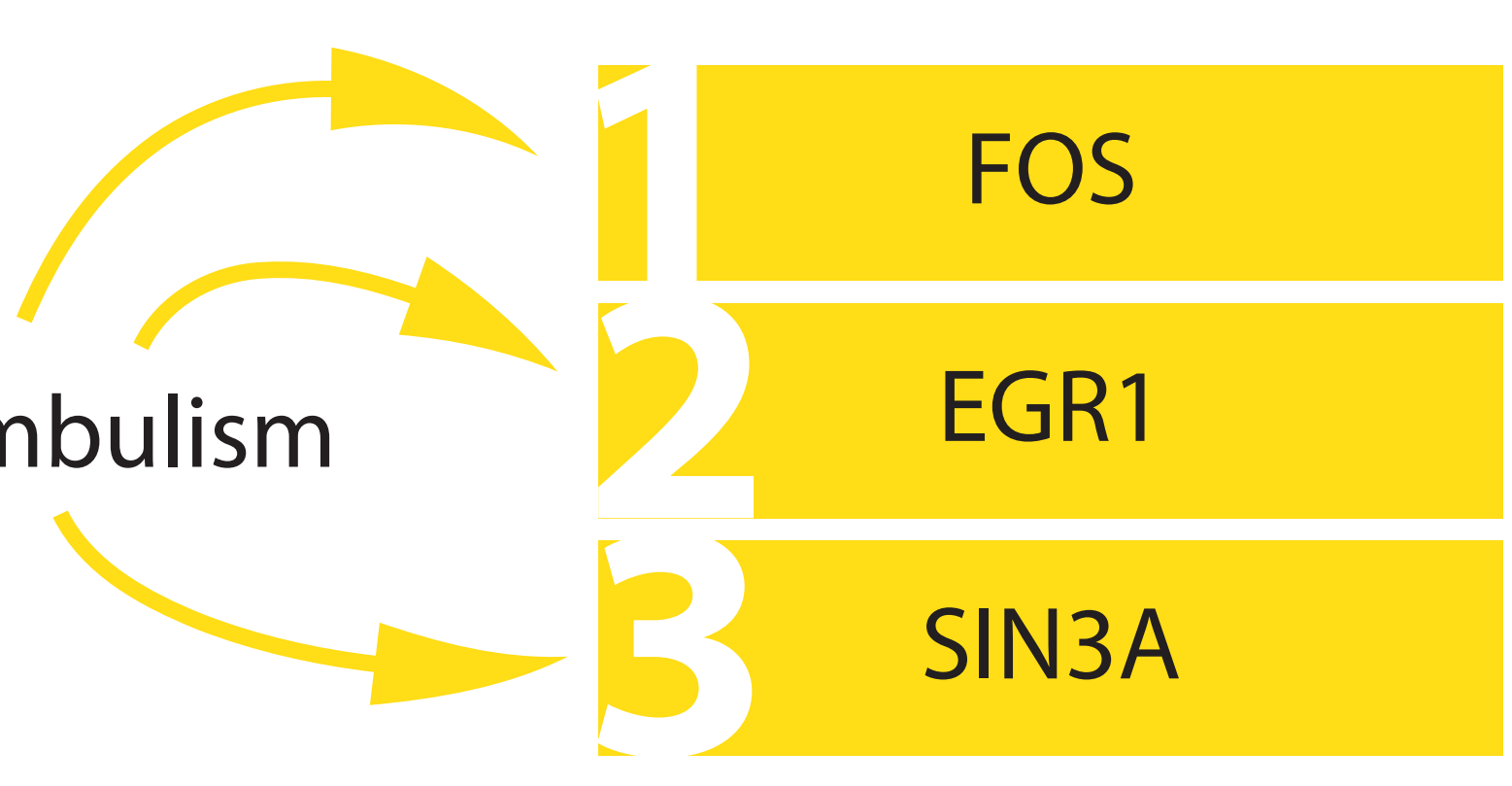
Towards improving the precision of a relation extraction system by processing negation and speculation

Vincent Van Asch, Roser Morante, Walter Daelemans
CLiPS - University of Antwerp

GOAL: Evaluating a direct implementation of a negation/speculation module.

Increasing precision of a relation extraction system by adding a negation/speculation filtering module.

CONTEXT



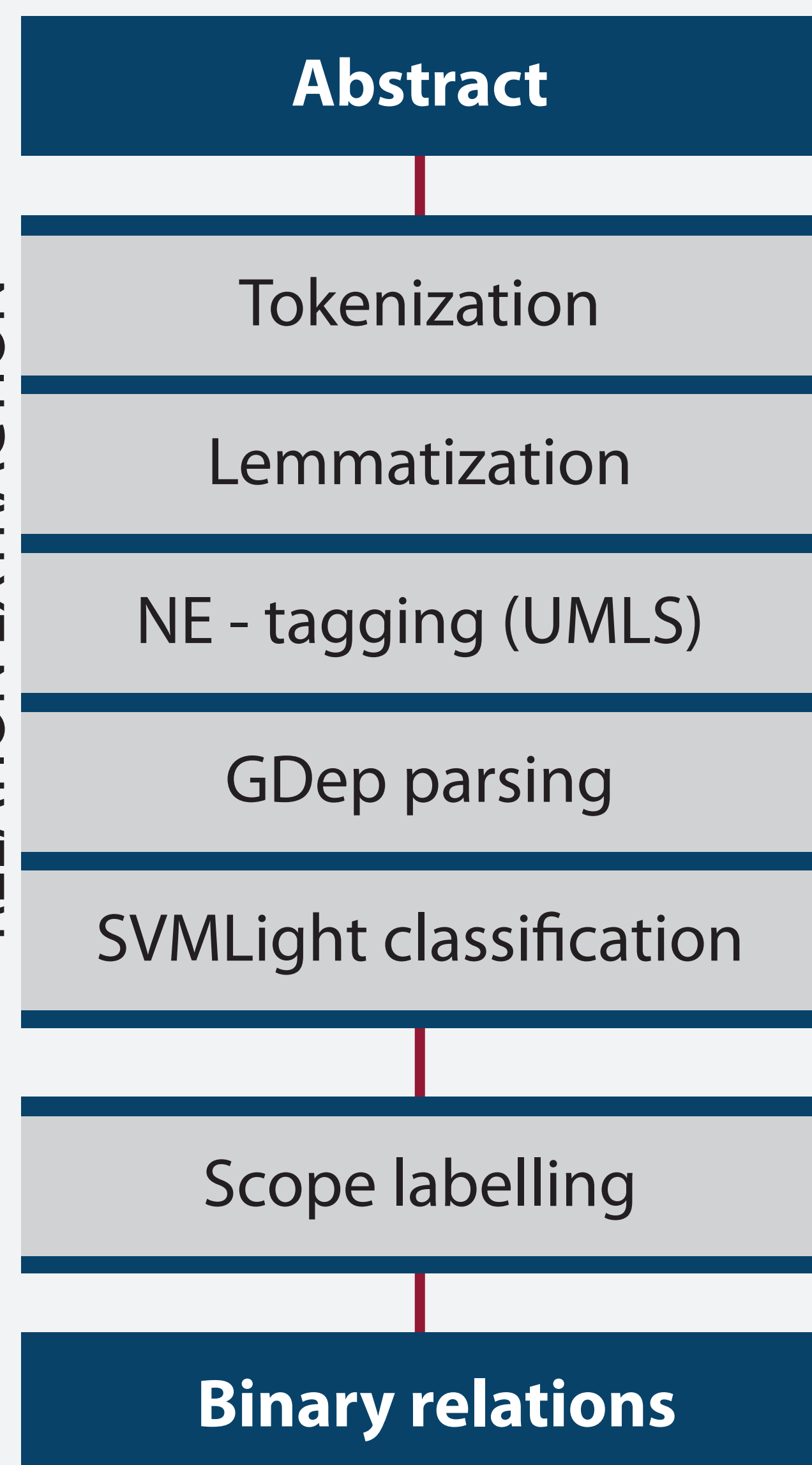
BIOGRAPH is a project that aims at putting forward a new methodology for text mining from heterogeneous information sources. The final goal of the project is to use text mining for finding new links between genes and phenotypes, and for gene prioritisation in order to discover non-obvious disease causing genes. BiographTA is the text analysis component of the project and is used to extract relations from biomedical abstracts, with a focus on precision.

CORPUS COLLECTION

Thus, **alpha-syntrophin** has an important role in synaps formation and in the organization of **utrophin**.

The binarized version of BioInfer corpus contains 2069 positive relations between 626 different named entities. By retrieving the full abstracts from PubMed, we have a full text corpus of 552 abstracts with annotated relations.

RELATION EXTRACTION



SYSTEM DESCRIPTION

The input for the relation extraction system are full abstracts. After some preprocessing steps, the actual relation extraction is done using SVMs. Every pair of named entities (NE) in a sentence is classified and accordingly gets the relation status or not.

The scope labelling module is optional and filters out negated relations. A relation is negated if both entities are under the scope of a negation and not under the scope of a speculation.

EXAMPLE OF FILTERING

[...] talin, but not vinculin or tubulin, appears to co-localize with actin microfilaments [...]

DISCUSSION

Adding a negation and speculation module increases the precision with 2.93%. Considering the fact that the goal of the BiographTA component is to extract relations with high precision, this is an encouraging result despite the lower F1-score.

The low recall is partly due to shortcomings of the NE-tagger. If we evaluate, only considering entities that the NE-tagger can handle, the system obtains a recall of 48.19% without scope labelling and 46.61% with scope labelling.

	precision	recall	F1-score
no scope labelling	69.61%	20.59%	31.78%
scope labelling	72.54%	19.91%	31.25%



More information about the BiographTA system is available at www.clips.ua.ac.be/BiographTA
Acknowledgements: The BIOGRAPH project was financed by the University of Antwerp - GOA.