IEEE
BIBM 2017

Named Entity Recognition in Functional Neuroimaging Literature

Asma Ben Abacha[1], Alba G. Seco de Herrera[2], Ke Wang[3], L. Rodney Long[1], Sameer Antani[1], Dina Demner-Fushman[1]

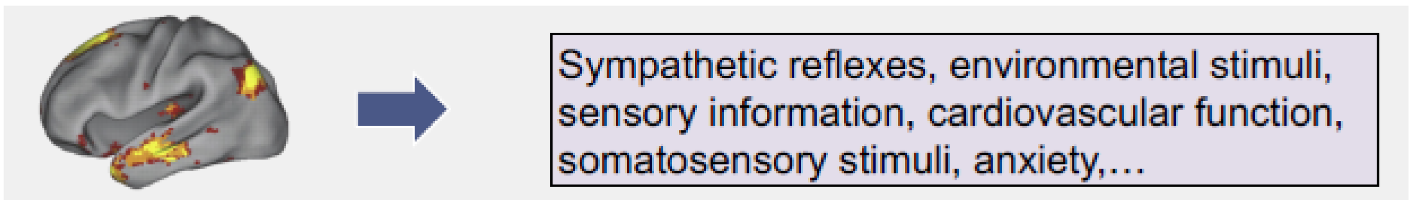
(1) U.S. National Library of Medicine, USA (2) University of Essex, UK (3) Swarthmore College, USA

Abstract

Human neuroimaging research aims to find mappings between brain activity and broad cognitive states. In particular, Functional Magnetic Resonance Imaging (fMRI) allows collecting information about activity in the brain in a non-invasive way. In this paper, we tackle the task of linking brain activity information from fMRI data with named entities expressed in functional neuroimaging literature. For the automatic extraction of those links, we focus on Named Entity Recognition (NER) and compare different methods to recognize relevant entities from fMRI literature. We selected 15 entity categories to describe cognitive states, anatomical areas, stimuli and responses. To cope with the lack of relevant training data, we proposed rule-based methods relying on noun-phrase detection and filtering. We also developed machine learning methods based on Conditional Random Fields (CRF) with morpho-syntactic and semantic features. We constructed a gold standard corpus to evaluate these different NER methods. A comparison of the obtained F_1 scores showed that the proposed approaches significantly outperform three state-of-the-art methods in open and specific domains with a best result of 78.79% F_1 score in exact span evaluation and 98.40% F_1 in inexact span evaluation.

I. Introduction

- Creating a detailed **map of brain function** requires an effective decoding of cognitive states from patterns of brain activity. Functional Magnetic Resonance Imaging (**fMRI**) has the capability of mapping brain activity to cognitive states.
- Around 65% of **fMRI research** study the properties of anatomical brain regions to explore functional localization, cognitive anatomy, or brain structures.
  - ✓ The **textual information** contained in the related publications forms a valuable resource for decoding cognitive states.
  - ✓ We focus on **recognizing named entities** that can be relevant to describe relationships between brain activity and a large number of broad cognitive states.



II. Data Construction

- We defined 15 named entity **categories**:
  - "Gross brain anatomy", "Functional neuroanatomy", "Brain function", "Body anatomy", "Body function", "Medical problem" and "Sensory stimuli or response" including eight sub-categories: "Gustation", "Visual", "Emotional", "Olfactory", "Auditory", "Somatosensory", "Abstract", and "Other".
- We developed **guidelines** for the annotation process and constructed a **gold standard corpus** of 52 neuroimaging abstracts manually double-annotated.

Figure 1 shows an annotated abstract using **Brat**.

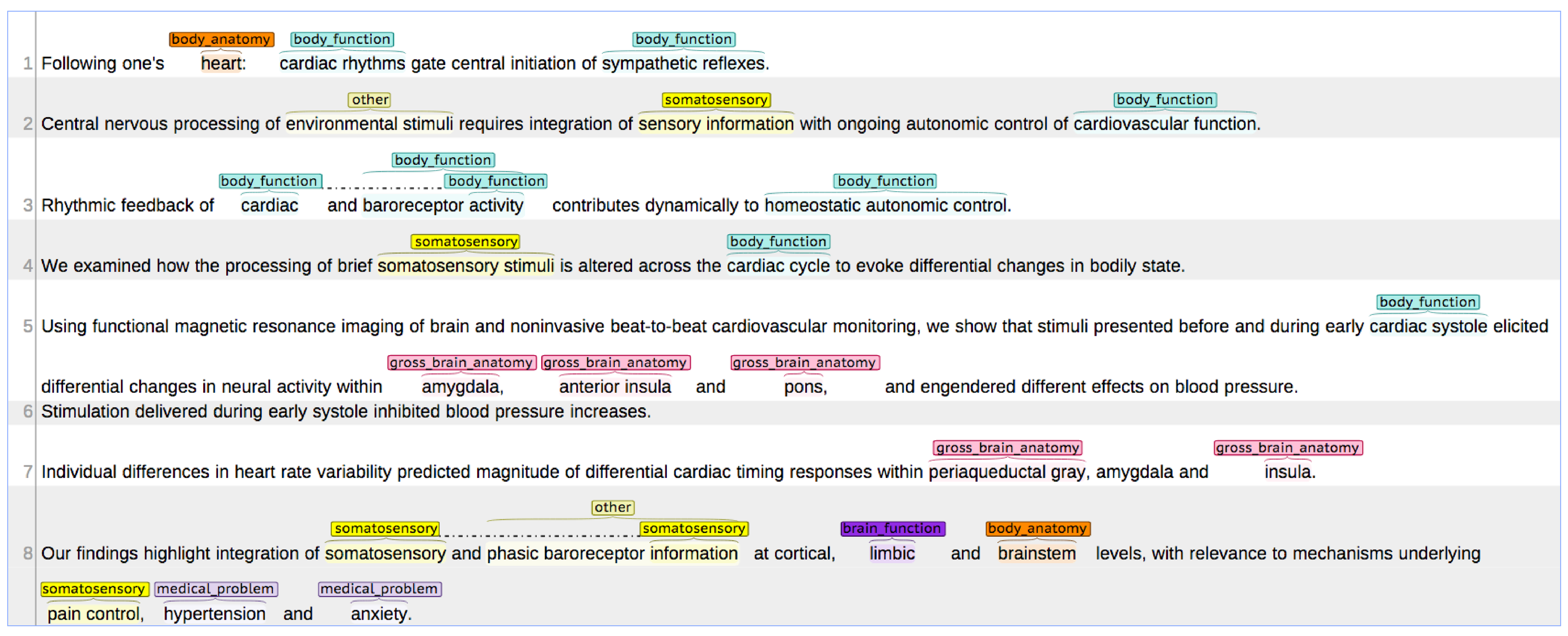


Table 1 shows the global Inter-Annotator Agreement (IAA) results. Precision (P), Recall (R) and F_1 score were computed for exact span matching and inexact matching (i.e. equal or overlapping).

Evaluation Criteria	Exact Span			Inexact Span		
	P	R	F_1	P	R	F_1
Span & Category	37.62	34.82	36.16	51.27	47.46	49.29
Entity Span Only	57.22	52.26	54.63	81.02	78.43	79.70

III. Methods

- We proposed two approaches for Named Entity Recognition (**NER**) from functional neuroimaging articles:
  - **Rule-based Methods:** To cope with the lack of training data for neuroimaging NER, as well as the inadequacy of existing biomedical NER systems targeting other categories (e.g. *Gene*, *Protein*, *Treatment*, *Chemical*), we proposed rule-based methods using noun-phrase detection combined with rules and filters designed for this specific task.
  - **Machine Learning Methods:** We developed supervised methods based on Conditional Random Fields (CRF), morpho-syntactic and semantic features (e.g. lemmas, POS tags, list of 7,062 diseases, gross brain anatomy terminology with 554 terms). CRF classifiers were trained on 91 manually annotated neuroimaging abstracts.
- We compared our methods to three **baselines**:
  - **Neurosynth**: a widely used platform for automatically synthesizing the results of different neuroimaging studies.
  - **DBPedia-KODA**: one of the best performing entity linking tools. We used the DBpedia implementation of KODA.
  - **MeSH®**: a controlled and structured vocabulary of medical topics provided and maintained by the U.S. National Library of Medicine.

IV. Results

Table 2 presents the results of Machine Learning (ML), Rule-based (RB), and Baseline (BL) methods.

Methods		Exact Matching			Inexact Matching		
		P	R	F_1	P	R	F_1
ML	CRF (with Semantic Features)	81.09	76.62	78.79	99.75	97.09	98.40
	CRF (without Semantic Features)	69.37	32.20	43.98	91.90	46.07	61.37
RB	TreeTagger + Rules	30.67	55.70	39.55	50.32	86.37	63.59
	Stanford Parser + Rules	23.92	61.40	34.43	51.02	93.44	66.00
BL	DBPedia-KODA	12.97	51.16	20.70	39.21	92.94	55.15
	MeSH®	4.39	3.84	4.09	12.37	11.80	12.08
	Neurosynth	3.08	19.65	5.3	32.12	98.11	48.40

V. Conclusions

- We studied NER for the automated understanding of brain anatomy and the brain cognitive functions expressed in publications related to fMRI experiments.
- Results showed that the proposed NER methods outperform state-of-the-art methods in open and specific domains.
- Using CRF and relevant semantic features, we achieve 78.79% F_1 score in exact evaluation and 98.40% F_1 in inexact evaluation.
- Our final goal is to develop a robust tool to map fMRI brain activations with relevant entities and thus, decode cognitive states from brain activity.
- This capability will allow using and integrating information contained in neuroscience publications at a large scale.

Contact

asma.benabacha@nih.gov
alba.garcia@essex.ac.uk