Disability annotation on documents from the biomedical domain

Alvarez Vecino, Pol pol.alvarez.vecino@est.fib.upc.edu

Advanced Human Language Technologies

Spring 2018

Introduction: Task

- **Annotate disabilities** and their textbfnegation on documents from the biomedical domain.
- Proposed as part of **IberEval 2018** workshop.
- Input documents are **short texts** with the disabilities and negations **tagged with XML**.
- Simple disability annotation:

```
... reliability of the MCA in Spanish to identify <dis>
   mild cognitive impairment</dis> (<dis>MCI</dis>)...
```

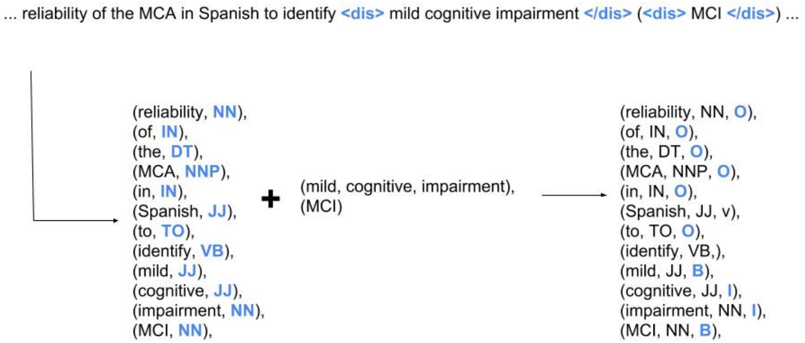
- Negated disability annotation:

```
... <scp> <neg>without</neg> <dis>dementia</dis> </scp>,
    significant differences were obtained in terms ...
```

Approach

- The approach to the problem was to solve it in two steps:
- **Disabilities:**
  - Convert all words into tuples: *(word, POS, IOB-tag)*
  - Use Conditional Random Fields (CRF) to predict the disability IOB-tags.
  - Convert IOB-tags to sentences/disabilities.
- **Negation:**
  - Feed each tuple sentence/disability to a negation software
  - Filter out the probable false positives
  - Convert input back again into XML tagged files.

Creating the training data



... reliability of the MCA in Spanish to identify <dis> mild cognitive impairment </dis> (<dis> MCI </dis>) ...

(reliability, NN),
(of, IN),
(the, DT),
(MCA, NNP),
(in, IN),
(Spanish, JJ),
(to, TO),
(identify, VB),
(mild, JJ),
(cognitive, JJ),
(impairment, NN),
(MCI, NN),

+

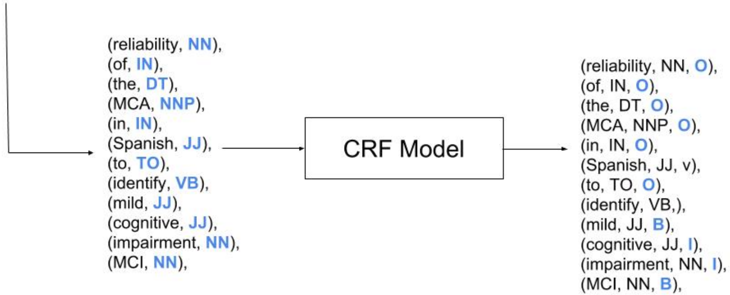
(mild, cognitive, impairment),
(MCI)

→

(reliability, NN, O),
(of, IN, O),
(the, DT, O),
(MCA, NNP, O),
(in, IN, O),
(Spanish, JJ, v),
(to, TO, O),
(identify, VB,),
(mild, JJ, B),
(cognitive, JJ, I),
(impairment, NN, I),
(MCI, NN, B),

Predicting the IOB-tags

... reliability of the MCA in Spanish to identify mild cognitive impairment (MCI) ...



(reliability, NN),
(of, IN),
(the, DT),
(MCA, NNP),
(in, IN),
(Spanish, JJ),
(to, TO),
(identify, VB),
(mild, JJ),
(cognitive, JJ),
(impairment, NN),
(MCI, NN),

CRF Model

(reliability, NN, O),
(of, IN, O),
(the, DT, O),
(MCA, NNP, O),
(in, IN, O),
(Spanish, JJ, v),
(to, TO, O),
(identify, VB,),
(mild, JJ, B),
(cognitive, JJ, I),
(impairment, NN, I),
(MCI, NN, B),

CRF model: Features (I)

Once the data in is $(word, pos, iob)$-format the next step is to decide which features are going to be used to train the CRF model.

Groups of features:

- 1. Derived from the **curent word**:
  - *word, pos, lemma, all-caps, strange-cap, contains-dash, contains-dot*
- 2. Features build from **entitites/acronyms lists**:
  - *inside-entities, is-acronym, position-in-entities, total-position-X*
- 3. Features from the **previous/next words**:
  - *prev-X-word, prev-X-pos, prev-X-lemma, next-X-word, next-X-pos, next-X-inside-entities*

CRF model: Features (II)

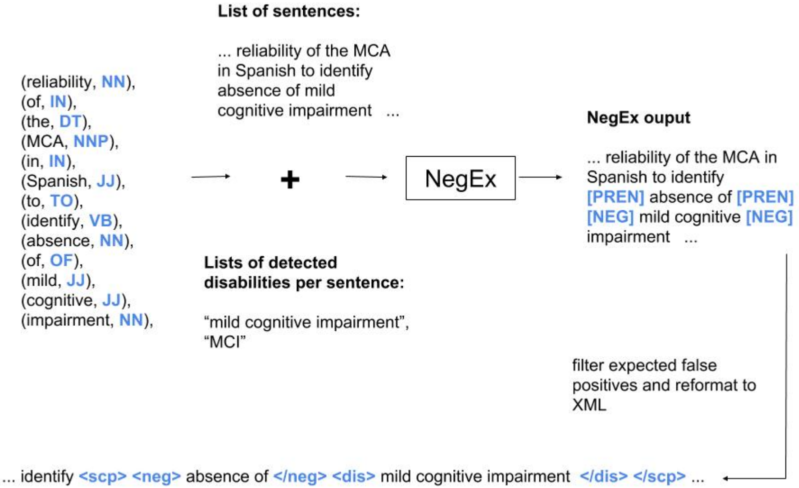
- 4. Features resulting from **concatenating a feature of the current word and one of the next/previous** one:
  - *prev1-word, prev1-pos, prev1-lemma*
  - *next1-word, next1-pos*
- 5. Features resulting from **concatenating a feature of the two previous/next** words:
  - *prev2-word, prev2-pos*
  - *next2-word, next2-pos*

CRF model: Training

- Feature Selection[1]:
  - Start with all groups activated and with all features per group
  - Deactivate a group and check if precision increases/decrease
  - If precision increases:
    - Reactivate the group
    - Deactivate each feature of the group and reactivate it only if precision decreases
  - If precision decreases just remove the group from the feature's set.
- Lists' Creation:
  - During model evaluation, entities/acronyms lists are created out of the training fold
  - Once the model is chosen, built the lists with all the training data.

[1] All validation results use 10-fold cross-validation

Negation Detection



(reliability, NN),
(of, IN),
(the, DT),
(MCA, NNP),
(in, IN),
(Spanish, JJ),
(to, TO),
(identify, VB),
(absence, NN),
(of, OF),
(mild, JJ),
(cognitive, JJ),
(impairment, NN),

List of sentences:

... reliability of the MCA in Spanish to identify absence of mild cognitive impairment ...

+

Lists of detected disabilities per sentence:

"mild cognitive impairment", "MCI"

NegEx ouput

... reliability of the MCA in Spanish to identify [PREN] absence of [PREN] [NEG] mild cognitive [NEG] impairment ...

filter expected false positives and reformat to XML

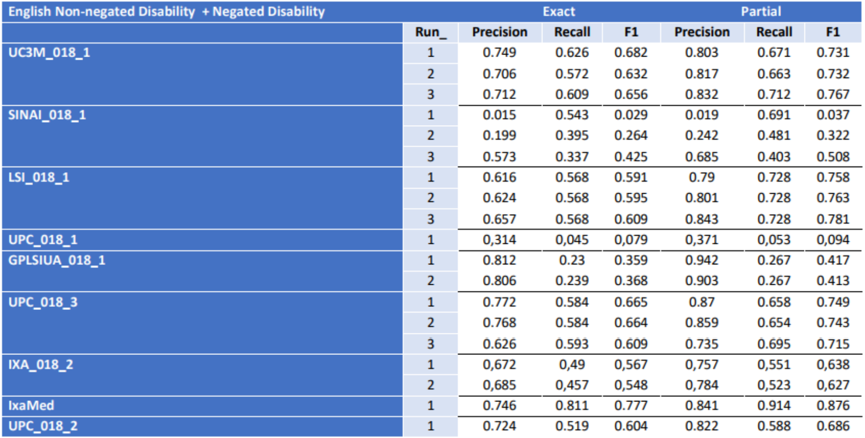
... identify <scp> <neg> absence of </neg> <dis> mild cognitive impairment </dis> </scp> ...

Technologies Used

- Tokenization and XML tag extraction:
  - NLTK tokenizer
  - LXML for XML tag extraction
- PoS tagging:
  - NLTK PoS Taggers
- CRF tagger:
  - NLTK Tagger interfaces
- Negation detector:
  - NegEx
- Precision evaluator: provided by DIANN organizers

English Results

The systems produces average results, ranking in the middle of the table for most metrics. In the overall English results, the system is 5th out of 9.

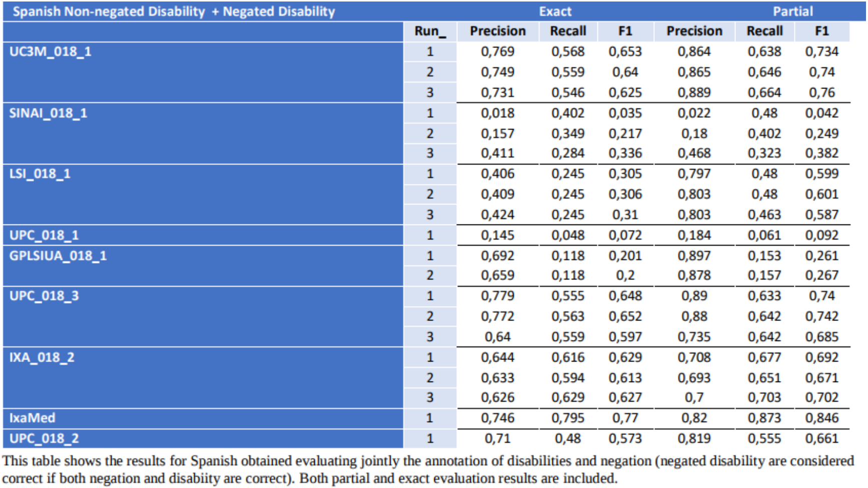


English Non-negated Disability + Negated Disability		Exact			Partial		
	Run_	Precision	Recall	F1	Precision	Recall	F1
UC3M_018_1	1	0.749	0.626	0.682	0.803	0.671	0.731
	2	0.706	0.572	0.632	0.817	0.663	0.732
	3	0.712	0.609	0.656	0.832	0.712	0.767
SINAI_018_1	1	0.015	0.543	0.029	0.019	0.691	0.037
	2	0.199	0.395	0.264	0.242	0.481	0.322
	3	0.573	0.337	0.425	0.685	0.403	0.508
LSI_018_1	1	0.616	0.568	0.591	0.79	0.728	0.758
	2	0.624	0.568	0.595	0.801	0.728	0.763
	3	0.657	0.568	0.609	0.843	0.728	0.781
UPC_018_1	1	0,314	0,045	0,079	0,371	0,053	0,094
GPLSIUA_018_1	1	0.812	0.23	0.359	0.942	0.267	0.417
	2	0.806	0.239	0.368	0.903	0.267	0.413
UPC_018_3	1	0.772	0.584	0.665	0.87	0.658	0.749
	2	0.768	0.584	0.664	0.859	0.654	0.743
	3	0.626	0.593	0.609	0.735	0.695	0.715
IXA_018_2	1	0,672	0,49	0,567	0,757	0,551	0,638
	2	0,685	0,457	0,548	0,784	0,523	0,627
IxaMed	1	0.746	0.811	0.777	0.841	0.914	0.876
UPC_018_2	1	0.724	0.519	0.604	0.822	0.588	0.686

This table shows the results for English obtained evaluating jointly the annotation of disabilities and negation (negated disability are considered correct if both negation and disabiity are correct). Both partial and exact evaluation results are included.

Spanish Results

In the overall Spanish results, the system is again 5th out of 9.



Spanish Non-negated Disability + Negated Disability		Exact			Partial		
	Run_	Precision	Recall	F1	Precision	Recall	F1
UC3M_018_1	1	0,769	0,568	0,653	0,864	0,638	0,734
	2	0,749	0,559	0,64	0,865	0,646	0,74
	3	0,731	0,546	0,625	0,889	0,664	0,76
SINAI_018_1	1	0,018	0,402	0,035	0,022	0,48	0,042
	2	0,157	0,349	0,217	0,18	0,402	0,249
	3	0,411	0,284	0,336	0,468	0,323	0,382
LSI_018_1	1	0,406	0,245	0,305	0,797	0,48	0,599
	2	0,409	0,245	0,306	0,803	0,48	0,601
	3	0,424	0,245	0,31	0,803	0,463	0,587
UPC_018_1	1	0,145	0,048	0,072	0,184	0,061	0,092
GPLSIUA_018_1	1	0,692	0,118	0,201	0,897	0,153	0,261
	2	0,659	0,118	0,2	0,878	0,157	0,267
UPC_018_3	1	0,779	0,555	0,648	0,89	0,633	0,74
	2	0,772	0,563	0,652	0,88	0,642	0,742
	3	0,64	0,559	0,597	0,735	0,642	0,685
IXA_018_2	1	0,644	0,616	0,629	0,708	0,677	0,692
	2	0,633	0,594	0,613	0,693	0,651	0,671
	3	0,626	0,629	0,627	0,7	0,703	0,702
IxaMed	1	0,746	0,795	0,77	0,82	0,873	0,846
UPC_018_2	1	0,71	0,48	0,573	0,819	0,555	0,661

This table shows the results for Spanish obtained evaluating jointly the annotation of disabilities and negation (negated disability are considered correct if both negation and disabiity are correct). Both partial and exact evaluation results are included.

Thank you.