Natural Language Analysis of Online Health Forums

Research Student
Abul Hasan

Supervisors
Mark Levene
David Weston

Department of Computer Science and Information Systems

Research Aims

Despite advances in concept extraction from free text, finding meaningful health related information from online patient forums still poses a significant challenge. We demonstrate a rule based natural language processing (NLP) system to extract structured information from posts found in such online health related forums by forming relationships between a drug/treatment and a symptom or side effect, including the polarity/sentiment of the patient. We termed this relationship as *disease triple* or simply a *triple*.

A motivating example

I take 600mg of gabapentin at bedtime, helps me shake and kick less; and a donepezil 10mg, settles me down allowing sleep. Clonazapam works great but I can't take the groggy, foggy head the next day.

Sentence or Sentence segment	Disease Triple
I take 600mg of gabapentin$_{drug}$ at bedtime, helps$_{+}$ me shake$_{symp}$ and$_{list}$ kick$_{symp}$ less$_{-}$;	(gabapentin,+,shake) (gabapentin,+,kick)
and$_{con}$ a donepezil$_{drug}$ 10mg, settles$_{+}$ me down allowing sleep$_{symp}$.	(donepezil,+,sleep)
Clonazapam$_{drug}$ works$_{+}$ great$_{intens}$	(Clonazapam,+,?)
but$_{con}$ I can't take$_{-}$ the groggy$_{side}$, foggy head$_{side}$ the next day.	(Clonazapam$_{drug\ anaphora}$,-,groggy) (Clonazapam$_{drug\ anaphora}$,-,foggy head)

Table 1. Triples extracted after processing the post. Subscripts denote different concept annotation.

Research Methodology

The overall methodology is schematically shown in Figure 1. The NLP pipeline splits the posts into sentences, tokenises the text and labels the tokens with their POS (parts-of-speech) tags. We disambiguate multiple concepts recognised from dictionary and ontologies by applying different linguistic rules, then split the sentences into segments and compute the polarity score of each resulting segment in stage 2 as shown in Figure 1. At stage 3, we first perform anaphora resolution and then form triples from each sentence or sentence segments.

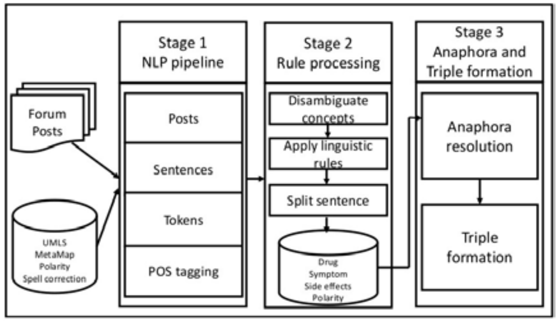

Figure 1. Text processing architecture

Experiment and Evaluation

We extracted 1056 user comments from discussion threads related to Parkinson's disease from PaitentsLikeMe (http://www.patientslikeme.com) website. 500 posts were used for training and 400 for testing the system. The remaining posts were used for the annotation validation. To evaluate our proposed approach the standard measures of accuracy, precision, recall and F_1 were used. Test results are shown in Table 2.

Concept	Accuracy	Precision	Recall	F_1 Score
Drug	90.71%	88.29%	95.14%	91.59%
Symptom	94.26%	84.08%	87.36%	85.69%
Side effects	98.42%	80.25%	93.53%	86.38%
Positive Polarity	86.44%	72.68%	94.42%	82.13%
Negative Polarity	87.08%	73.57%	88.52%	80.35%
Triple1	73.93%	71.11%	96.02%	81.71%
Triple 2	74.47%	71.31%	96.81%	82.13%

Table 2. Test results for 400 posts

Concluding Remarks and Future Works

We have developed a strong baseline system, achieving an F_1 score of over 80%, in identifying disease triples. Our next goal is to transfer the knowledge gained in our system to discover triples in other patient forums. In order to achieve this we anticipate the use of machine learning methods, which can adapt to different usage of language than the PatientsLikeMe forum we have concentrated on.