SALÃO DE INICIAÇÃO CIENTÍFICA XXIX SIC

UFRGS PROPESQ

Salão UFRGS 2017

múltipla UNIVERSIDADE inovadora inspiradora

Evento	Salão UFRGS 2017: SIC - XXIX SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS
Ano	2017
Local	Campus do Vale
Título	Detectando condições clínicas com processamento de linguagem natural
Autor	FELIPE SOARES FAGUNDES PAULA
Orientador	ALINE VILLAVICENCIO

Detectando condições clínicas com processamento de linguagem natural.[1]

Autor: Felipe Paula, *Instituto de Informática* - UFRGS
Orientadora: Aline Villavicencio, *Instituto de Informática* - UFRGS

Neste trabalho propomos uma abordagem automática para detecção de doenças neuropsiquiátricas através do processamento de dados de teste de fluência verbal semântica (TFV). Esses testes estão em várias baterias de avaliação neuropsicológicas e são chave na identificação de demências como Alzheimer e Déficit Cognitivo Leve. Nos TFVs, é pedido para o paciente falar o maior número de palavras pertencentes a uma determinada categoria, como animais, em um tempo pré-determinado, como 60 segundos (ex. cachorro, gato, boi, porco, leão, zebra, girafa, …). Subgrupos semânticos, como *pets*, "cachorro e gato", animais de fazenda, "boi e porco" e animais africanos, "zebra e girafa" são chamados de *clusters* e a alternância entre diferentes grupos chamamos de *switches* e eles estão relacionados ao funcionamento da memória semântica e às funções executivas dos indivíduos. Logo, alterações nos *clusters* e *switches* podem ser boas pistas na predição dessas doenças. Para a detecção dessas anomalias, propomos o uso de medidas de similaridade semântica para detectar quais pares de palavras estão menos fortemente associados e, por conseguinte, provavelmente são switches. Transformamos os dados dos TFV em representações num espaço semântico onde se assume que a proximidade no espaço corresponde a proximidade semântica num mesmo cluster. Isso é feito pegando os pares de palavras em sequência e aplicando uma medida de similaridade entre elas. Para chegarmos a essa representação, usamos diferentes modelos, como Skipgram, GloVe e LexVec, construídos automaticamente a partir de textos da Wikipedia. Além disso, usamos como baseline uma medida de força de associação baseada no ppmi (*positive pointwise mutual information*) da coocorrência de palavras no mesmo corpus. Os switches podem ser detectados, por exemplo, quando os pares de palavras estão abaixo da média de similaridade de determinado teste. Depois de detectar os *switches* e os *clusters*, extraímos atributos para construir um classificador como: o número de palavras, o número de switches, tamanho médio de cluster, etc. Utilizamos classificadores *random forest* para separar os grupos clínicos. Para comparar a performance das nossas heurísticas, também utilizamos a informação de *cluster* e *switch* informadas por um especialista - que representa um gold standard. Na separação entre o grupo com Alzheimer e grupo de controle, obtivemos um AUC de 0.89 (desvio = 0.02) com o classificador baseado em heurísticas e AUC de 0.87 (desvio = 0.03) com o classificador treinado com o gold standard. Isto é, nosso método apresenta melhor performance, $t(99) = 6.12$ ($p < 0.05$). Adicionalmente, obtivemos um resultado melhor $t(99) = 12.68$ ($p < 0.05$) que o método previamente utilizado para esse conjunto de dados, de AUC de 0.85 (desvio = 0.02). Esse resultado é promissor e indica que o nosso método pode ajudar a construir ferramentas automáticas para auxiliar no diagnóstico de doenças neuropsiquiátricas. Outra conclusão interessante é que as heurísticas propostas não apenas conseguem capturar o comportamento da dinâmica de cluster e switch, mas também podem possuir mais informação necessária para o diagnóstico dessas doenças. Esse fato pode revelar que essas heurísticas também podem ser complementares às análises de cluster e switch realizadas por especialistas, que muitas vezes se atém a um classificação de grupos semânticos baseada em uma taxonomia bastante rígida. Gostaríamos de investigar como essas dinâmicas se traduzem em outros idiomas, como o inglês, principalmente por existirem outros recursos que poderíamos utilizar, como por exemplo, a WordNet. Também gostaríamos de analisar mais de perto as diferenças entre os *clusters* e *switches* propostos por especialistas e os sugeridos pelas nossas heurísticas.

1. Trabalho aceito para apresentação oral no 2017 WiNLP workshop - *Detecting clinical conditions with Distributional Semantic Models*, Felipe Paula, Rodrigo Wilkens, Marco Idiart and Aline Villavicencio