https://telemedicina.unifesp.br/pub/Events/2013-05%20-%20WWW2013/www2013/www2013.org/companion/p103.pdf

Método em Zhang e Zhu, 2013

Dado um corpus textual D (conjunto de reviews)

Passo 1: Construir uma matriz quadrada C com a coocorrência de todas palavras a nível de oração. (sem artigos, preposições, etc.)

Passo 2: Montar M, matriz de modificação, que relaciona opiniões e features. Sua construção é iterativa: se escolhe um conjunto inicial de opiniões O_0, para cada opinião nesse conjunto, escaneia-se as reviews R em D e se extrai as features (explícitas) correspondentes, gramaticalmente anotando em M. Depois, faz o mesmo com F, as feats encontradas; e se repete iterativamente

Passo 3: São selecionadas as características candidatas F_c para uma determinada review R_i com conjunto de palavras de opinião O definido. Essas features candidatas são as features em M que podem ser modificadas por O.

Passo 4 - A EXTRAÇÃO: Dada uma revisão R, sem features explícitas, com conjunto W (de tamanho w) de palavras $W = \{w_1, w_2, \ldots, w_w\}$, se obtem $W_- = W - (W \cap F_c)$, com F_c o conjunto de features candidatas do passo 3. Em seguida, é construída a matriz S $|F_c| \times |W_-|$ (submatriz de C do passo 1). As linhas da matriz são as reduções candidatas, enquanto as colunas são as palavras em R.

Segundo S, é possível obter as probabilidades de coocorrência. Supõe-se m_m palavras em D, com m_a aparições de w_a e m_b aparições de w_b. w_a e w_b coocorrem m_{ab}* vezes na mesma oração. Tem-se então:

$$P(w_a \mid w_b) = \frac{P(w_a, w_b)}{P(w_b)} = \frac{m_{ab}/m_m}{m_b/m_m} = \frac{m_{ab}}{m_b} \qquad (1)$$

Por fim, então, para cada feature $f_i \in F$ se calcula o valor médio da probabilidade condicional acima com todas $w_i \in W_-$

* m_c no artigo original

$$T(f_i) = \frac{1}{\omega} \sum_{j=1}^{\omega} P(f_i | \omega_j) \qquad (2)$$

a feature escolhida é $\arg\max T(f_i)$

Análise

↳ Método fundamentada em associação por coocorrência;

↳ Bons resultados nos bases chimeras testadas pelos autores;

↳ Não-supervisionado: não requer anotação humana, não divide em treino/teste;

↳ Requer volume de dados grande o suficiente para fornecer boas aproximações dos índices de coocorrência;

↳ É feita a forte presunção de que o conjunto de features explícitas em D contém todas features implícitas. Além disso, se presume que uma feature vem acompanhada das

mesmas palavras, independentemente de estar explícito ou implícito.

Mudança proposta por Schouten e Frasincar, 2014

↳ O conjunto de features candidatas F_c é composto por features implícitas anotadas, resultando em medidas de coocorrência diretas entre as palavras e features implícitas;

↳ Portanto, agora o algoritmo é supervisionado

↳ A versão revisada do algoritmo tem melhor desempenho nos testes dos autores.

↳ Foram testados modelos com todas combinações possíveis de inclusão ou não de substantivos, verbos, adjetivos e advérbios, obtendo melhores resultados com a combinação substantivos, verbos e adjetivos.

Indo além...

Foi dito acima que alguns processos são

feitos "gramaticalmente". Como, por exemplo, se sabe que opinião se refere a qual feature explícita? Como são identificados os substantivos, núcleos, etc.?

Para isso, são usados softwares específicos, os Natural Language Parsers. É possível encontrar alguns de licença aberta para variadas linguagens.

SpaCy, NLTK: python

Open NLP: R