VARRA: um serviço para a Validação, Avaliação e Revisão de Relações semânticas no AC/DC

Claudia Freitas, Diana Santos, Hugo Gonçalo Oliveira & Violeta Quental

A importância dos recursos lexicais – léxicos, ontologias, tesauros - para sistemas que lidam com o processamento computacional da língua é cada vez mais reconhecida, assim como as dificuldades inerentes a sua elaboração. De um lado, metodologias baseadas na extração automática de relações semânticas entre palavras a partir de corpus têm, na forma de avaliação/validação das relações, o seu ponto fraco. De outro, a elaboração manual, por sua vez, se garante uma precisão dos resultados, depende de um processo altamente custoso, que se busca normalmente evitar.

Nesse contexto, apresentamos o VARRA, um sistema que, por meio da ocorrência de pares de palavras em textos, pretende auxiliar a validação de relações semânticas entre essas palavras. Cada relação é representada como uma tripla *palavra1_relacao_palavra2* (por exemplo "*mentira_sinônimo_ilusão*").

O VARRA foi desenvolvido no âmbito do projeto AC/DC (uma interface comum para acesso e disponibilização de corpora em português), de forma a obter, de maneira mais objetiva, julgamentos de falantes nativos quanto às relações semânticas em questão, buscando validá-las a partir do uso das palavras em contextos autênticos, representados por frases dos corpora do projeto AC/DC.

Buscamos, com o VARRA, construir uma base confiável de julgamentos sobre uma dada relação semântica entre pares de palavras, assim como criar um procedimento de avaliação parecido com a interpretação humana (em oposição à validação de relações entre palavras fora de contexto). Ao invés de perguntarmos, por exemplo, se "mentira é sinônimo de ilusão", perguntamos se um dado contexto/frase ilustra a relação de sinonímia entre *mentira* e *ilusão*. Caberá ao avaliador julgar, auxiliado pelo contexto, se a relação é possível ou não. Nos trechos abaixo, por exemplo, as respostas seriam SIM e NÃO, respectivamente.

> *Mudança maior, porém, vem do novo presidente do Supremo Tribunal Federal, ministro Sepúlveda Pertence, que afirmou: ` Desde que se superou a* ***mentira*** *de que um juiz, particularmente um juiz constitucional, é um puro técnico capaz de extrair uma norma supostamente de um único sentido válido de um fato, desde que essa* ***ilusão*** *foi desfeita, a verdade é que o juiz é um homem, enquanto cidadão, com crenças, convicções, tendências conscientes e inconscientes .* (Chave, AC/DC)

> *Há muitas acções que respeitadas à luz da liberdade podem não o ser quando desencadeiam a desordem, a mistificação, a* ***mentira****, a* ***ilusão****, a feitiçaria, o roubo, disse o prelado, acrescentando no entanto que como bispo da Igreja Católica, a sua função era a de impedir e lutar contra toda e qualquer caça às bruxas .* (Chave, AC/DC)

Como sabemos que, nem sempre, e principalmente em termos de relação semântica, SIM/NAO são respostas suficientes, para os casos em que o texto não ilustra as relações é possível ainda classificá-las como

(i) O contexto é insuficiente para validar a relação, embora seja, de alguma maneira, compatível com ela:

*A primeira vez que vi foi chocante, surpreendente, pelo jogo que faz entre verdade e **mentira**, pela discussão sobre a **ilusão** e seu poder, sobre a mistificação .* (Chave, AC/DC)

(ii) O texto não valida a relação, e é completamente não relacionado:

*Esta **ilusão** de ótica é tão interessante que até parece de **mentira**.*
(Frase extraída da internet)

(iii) O texto não valida a relação; pelo contrário, invalida-a:

*Não era uma **mentira**, era **ilusão**, o cinema possui esse poder de criar ilusão*. (Chave, AC/DC)

(iv) Não se pode obter qualquer conclusão a partir da frase exemplo

*Distingue, nas promessas que lhe fazem, o que é **mentira** ou **ilusão** .*
(Chave, AC/DC)

Além da validação de relações semânticas já previstas (sinonímia, hiperonímia, parte_de; causador_de, entre outros), O VARRA permite também a exploração e identificação de padrões léxico-sintáticos capazes de expressar relações semânticas entre pares de palavras. Para tanto, basta efetuar uma busca pelas palavras (selecionando a relação "qualquer relação") e verificar se é possível identificar algum padrão nos resultados das ocorrências. A busca pelos pares *belo/feio*, por exemplo, pode indicar, a partir dos contextos em que essas palavras ocorrem, padrões para a expressão da relação de antonímia.

Para a expressão de padrões novos, que nao estão na base do VARRA, uma grande funcionalidade é a possibilidade de considerar, na expressão de busca, informações linguísticas como classe de palavras e função sintática, o que permite uma procura altamente refinada por padrões em um corpus. Isto é possível porque, como mencionado, o VARRA tem como base os corpora do projeto AC/DC, que previamente analisados pelo analisador morfossintático PALAVRAS.

O objetivo deste trabalho é apresentar o sistema VARRA, bem como as possibilidades que este oferece para a investigação em semântica computacional e lexicografia.