ENEPEX
ENCONTRO DE ENSINO, PESQUISA E EXTENSÃO
12º ENEPE UFGD - 9º EPEX UEMS
A INTERNACIONALIZAÇÃO DA UNIVERSIDADE E O FORTALECIMENTO DO ENSINO

# CONVERSÃO AUTOMÁTICA, PARA RELAÇÕES BINÁRIAS, DE RELAÇÕES DE ASSOCIAÇÃO EXTRAÍDAS DE FORMA AUTOMATIZADA

BIASSIO, Rafael Oliveira[1] (rafaelobiassio@gmail.com); BATISTA Jr, Joinvile[2] (joinvile@ufgd.edu.br);
[1]PIVIC e [2] Professor do Curso de Sistemas de Informação da UFGD – Dourados.

## INTRODUÇÃO

Ontologias são construídas a partir de relações binárias (argumento 1 -- relação -- argumento 2). Os argumentos são majoritariamente associados a uma frase nominal (que não possui verbos e que tem substantivos como foco de sua semântica) e a relação é associada a uma frase verbal (que tem verbos como foco de sua semântica).

As extrações geradas pelo extrator aberto disponível OpenIE 4 [1], a partir de sentenças mais complexas, contém, na maioria das vezes, frases verbais fora da relação: (a) no argumento 1; (b) no argumento 2; ou (c) em trechos que foram inadequadamente descartados nas extrações geradas.

O objetivo deste trabalho é a remoção de frases verbais de argumentos e de trechos descartados, os quais estão sendo referenciados como argumentos complexos e trechos descartados complexos.

## MATERIAIS E MÉTODOS

A metodologia utilizada neste trabalho foi baseada nas seguintes etapas: (a) caracterização das situações envolvendo argumentos 1 complexos; (b) especificação e implementação de padrões para convertê-los em argumentos simples, com a conversão da frase verbal em um frase substantiva; (c) caracterização das situações envolvendo argumentos 2 e trechos descartados complexos, para os quais foi necessário especificar regras de conversão para gerar novamente as extrações de relações a partir da sentença original, para evitar a perda da semântica da sentença original; (d) geração das regras de conversão a partir das situações identificadas; e (e) implementação de um conversor genérico, baseado em regras definidas em XML para realizar as extrações, de forma que os argumentos não utilizassem frases verbais.

O corpus utilizado, para especificação, implementação e teste do protótipo, foi a seção “Food Choices” de um livro sobre nutrição básica, cujo texto expressa motivações sociais, religiosas e pessoais para escolha de alimentos, acarretando maior de dificuldade na extração de relações.

Para identificação de argumentos e trechos descartados complexos, bem como para o desenvolvimento de padrões de conversão de argumentos 1 complexos, foram utilizadas as seguintes ferramentas: (a) o extrator OpenIE 4 [1]; e (b) uma ferramenta desenvolvida anteriormente [2], para geração de padrões de representação das extrações geradas pelo OpenIE 4.

## RESULTADOS E DISCUSSÃO

A proposta inicial previa a transformação dos argumentos e dos trechos descartados complexos para construir relações binárias, de forma que frases verbais fossem utilizadas somente na relação da tripla que compõe a relação binária. Esta estratégia foi possível para o argumento 1 de uma dada relação. No entanto, no caso do argumento 2 ou de trecho descartado, não foi possível simplesmente convertê-los. Neste caso, foi necessária a definição de padrões de extração considerando a sentença como um todo, de forma a não descaracterizar a semântica das relações binárias geradas, do ponto de vista da semântica da sentença original.

Para a situação do argumento 1 complexo foram identificados seis padrões distintos.

Para caracterizar a necessidade de conversão da sentença como um todo, no caso de um argumento 2 complexo, são ilustrados:
sentença 13 do corpus: “Although most people realize that their food habits affect their health, they often choose foods for other reasons.”;
extrações geradas pelo OpenIE 4: (1) they -- choose -- foods for other reasons - often; (2) most people -- realize -- that their food habits affect their health;
extrações sugeridas para eliminar os argumentos 2 complexos sem perdar a semântica da sentença original: (1) Although:: most people -- realize -- that; (2) that:: their food habits -- affect -- their health; (3) they -- often choose -- foods - for other reasons;
antecedente da regra de extração especificada para gerar essas três extrações: SBAR NP1 VP1 |that| NP2 VP2 NP3 |,| NP4 @VP3 X1;
consequentes da regra: (1) SBAR:: NP1 -- VP1 -- |that|; (2) |that|:: NP2 -- VP2 -- NP3; e (3) NP4 -- @VP3 -- X1.

Foi desenvolvido um protótipo genérico capaz de tratar padrões de conversão representados em XML, para potenciar a prototipagem de novos padrões sem a necessidade de alterar o código do gerador do extrator de relações binárias. O padrão de conversão foi concebido com uma regra composta de um antecedente e de consequentes, de forma que quanto o texto de uma dada sentença combina com o antecedente do regra de conversão, os seus consequentes são aplicados para gerar as relações binárias desejadas. No corpus utilizado, foram especificados padrões de conversão para as sentenças com argumentos 2 complexos ou com trechos descartados complexos.

## CONCLUSÕES

O desenvolvimento de um conversor que trata genericamente regras de conversão especificadas em XML e gera as relações binárias, agiliza substancialmente a prototipagem de regras de conversão para a extração das relações binárias.

Pelo fato das regras estarem sendo definidas com base nas sentenças a serem convertidas, o resultado é obviamente muito superior ao gerado pelo OpenIE 4. No entanto, o OpenIE realiza a extração de forma totalmente automática, embora gere muitas extrações falhas e descarte trechos que não deveriam ser descartados. A notação especificada para representar as regras de extração, viabiliza a representação de regras genéricas devido a: (a) utilização preferencial de categorias de frases em vez de léxicos ou textos originais; (b) a utilização de encadeamentos de categorias de frases denotadas por @NP e @VP; e (c) a representação de trechos genéricos da sentença, denotados por “X”.

## REFERÊNCIAS

[1] ETZIONI, Oren; MAUSAM, Mausam; SCHMITZ, Michael. Open IE 4 Software (C) 2011-2012, University of Washington. US patent number 7,877,343 and 12/970,155 patent pending.

[2] DIUNISIO, Mateus Trindade; BATISTA Jr, Joinvile. Seleção Automática de Extrações de Relações de Associação Geradas de Forma Automatizada a partir de Texto Abordando um Tema de Nutrição. ENEPEX 2017.

Realização: UFGD Universidade Federal da Grande Dourados; UEMS Universidade Estadual de Mato Grosso do Sul
Parceiros: CAPES; CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico