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

EXTENSÃO DOS PADRÕES DE CLASSIFICAÇÃO COM PADRÕES DE EXTRAÇÃO DE RELAÇÕES VINCULADAS ÀS SENTENÇAS CLASSIFICADAS

COUTINHO, Jeovano de Oliveira[1] (jeovanocoutinho@gmail.com); BATISTA Jr, Joinvile[2] (joinvile@ufgd.edu.br);
[1]Bolsista PIBIC e [2] Professor do Curso de Sistemas de Informação da UFGD – Dourados.

INTRODUÇÃO

A classificação das sentenças, utilizando filtros baseados em padrões, foi desenvolvida em um trabalho anterior [1], para ser utilizada na construção de uma ontologia para caracterizar o papel de cada sentença relevente no artigo técnico como, por exemplo: objetivos, contribuições, principais resultados, vantagens, desvantagens.

O objetivo deste trabalho é o de agregar padrões de extração de relações, aos padrões de classificação, para que seja possível construir uma ontologia para representar artigos técnicos em um dada área do conhecimento, dado que a ontologia interliga entidades através de relações.

MATERIAIS E MÉTODOS

A metodologia utilizada neste trabalho foi baseada nas seguintes etapas: (a) escolha de padrões de classificação representativos, para selecionar as sentenças do corpus (texto de entrada) a ser utilizado; (b) extrações das sentenças selecionadas, com base nos padrões de classificação; (c) definição de uma notação para representação de padrões de extração; e (d) implementação e teste de uma ferramenta para gerar a extração das relações a partir dos padrões de extração.

Os padrões de extração de relações foram direcionados pelos padrões de classificação, para alcançar o objetivo de poder gerar relações com foco na classificação de artigos técnicos. O corpus utilizado neste trabalho foi composto de sentenças selecionadas do artigo técnico que havia sido utilizado para gerar os padrões de classificação [2].

RESULTADOS E DISCUSSÃO

Para caracterizar o papel dos padrões de classificação é ilustrada uma sentença, seu padrão de classificação, o padrão de extração gerado e as extrações resultantes.

A sentença utilizada como ilustração é:

- Search engines retrieve and rank potentially relevant documents for human perusal, but do not extract facts, assess confidence, or fuse information from multiple documents.

O padrão de classificação que havia sido gerado para esta sentença é o seguinte:

- Limitation: % $1 [|and| $2] % |, but do not| $3 %

  $1 = {Present}

  $2 = {Present}

  $3 = {Infinitive}

Em função do padrão de classificação, “but do not” passa a ser o elemento ligador entre as extrações geradas:

- Search engines -- retrieve and rank -- potentially relevant documents - for human perusal
- but:: Search engines -- do not extract -- facts - from multiple documents
- but:: Search engines -- do not assess -- confidence - from multiple documents
- but:: Search engines -- do not fuse -- information - from multiple documents

Para concluir a ilustração sentença, o padrão de extração pode ser representado pela seguinte regra:

- Antecedente

  NP1 VP1 @NP1 |, but do not| VP2 NP2 [|,| VP3 NP3 |, or| VP4 @NP2]

- Consequentes

  NP1 -- VP1 -- @NP1

  |but|:: NP1 -- |do not| VP2 -- NP2

  |but|:: NP1 -- |do not| VP3 -- NP3

  |but|:: NP1 -- |do not| VP4 -- @NP2

Neste caso, a conjunção coordenada “but” não pertence às extrações, mas está sendo salva na representação das extrações como um modificador da relação, servindo assim como um elemento de ligação entre a primeira extração e as demais. Desta forma a semântica da sentença original não é perdida. Em função da sequência opcional empregada no antecedente da regra, o terceiro e o quarto consequentes da regra somente serão utilizados se os elementos previstos na sequencia opcional forem encontrados na sentença original.

Foram definidos padrões de extração para várias sentenças selecionadas do artigo técnico escolhido, de forma a especificar uma sintaxe genérica para a geração dos padrões de extração, associados aos padrões de classificação. Foi então prototipada uma ferramenta para automatizar a geração das extrações a partir da representação do padrões de extração definidos em XML.

CONCLUSÕES

A geração de extrações, a partir de padrões de classificação definidos previamente, é o passo essencial para a construção de um ontologia para suportar prospecção tecnológica a partir de artigos técnicos, dado que provê as relações binárias e classificação do papel das relações no contexto do artigo.

A utilização desta ferramenta simplifica significativamente a geração de um conjunto maior de padrões de extração, dado que o acréscimo de um novo padrão de extração não implica na alteração do código da ferramenta.

A partir do resultado deste trabalho, poderá ser gerado um conjunto significativo de padrões de extração, a partir de padrões de classificação, para servir como base para a especificação de um processo automático de geração de padrões de extração.

REFERÊNCIAS

[1] LIMA, Herlon Augusto Aguiar; BATISTA Jr, Joinvile. Prototipagem da classificação automática dos papéis de trechos das sentenças para prospecção tecnológica a partir de um artigo técnico. ENEPEX, 2016.

[2] ETZIONI, Oren; CAFARELLA, Michael; DOWNEY, Doug; KOK, Stanley; POPESCU, Ana-Maria; SHAKED, Tal; SODERLAND, Stephen; WELD, Daniel Sabey; YATES, Alexander. Web-scale information extraction in KnowItAll (preliminary results). Proceedings of the 13th International Conference on the World Wide Web, 2004.

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