ENEPEX
ENCONTRO DE ENSINO, PESQUISA E EXTENSÃO
8° ENEPE UFGD • 5° EPEX UEMS

DESENVOLVIMENTO DE PROTÓTIPO PARA CRIAÇÃO AUTOMÁTICA DE ONTOLOGIA NÚCLEO A PARTIR DA *WORDNET*

Joinvile Batista Junior[1]; Reinaldo Costa da Silva[2];
UFGD-FACET, C. Postal 533, 79804-970 Dourados-MS, E-mail: joinvile@ufgd.edu.br
[1] Professor do curso de Sistemas de Informação. [2] Aluno voluntário de Iniciação Científica.

RESUMO

Ontologias representam conhecimento em um dado domínio. Pesquisas semânticas utilizam o sentido das relações entre conceitos definidas em ontologias, indo além da pesquisa envolvendo palavras chaves. Ontologias têm sido construídas a partir de relações binárias, compostas de duas entidades relacionadas através de verbos, extraídas de sentenças em linguagem natural. A *WordNet* é uma ontologia, construída manualmente, que tem sido extensamente utilizada para pesquisas em processamento de linguagem natural, e que suporta informações sobre a semântica de substantivos, adjetivos e verbos da língua inglesa. Para cada substantivo, a *WordNet* provê, entre várias informações, os vários sentidos associados ao substantivo e um texto explicativo associado a cada sentido do substantivo em questão. O objetivo deste trabalho é o desenvolvimento de um protótipo para extração de relações binárias, para posterior construção de uma ontologia núcleo, a partir dos textos explicativos da *WordNet* relativos a conceitos oriundos de sentenças extraídas de um corpus correspondente a uma fonte fidedigna: um livro de referência para o ensino de nutrição básica. A metodologia utilizada incorpora: (a) a seleção automática de conceitos relevantes do domínio de nutrição; (b) a seleção automática do sentido, associado a cada conceito na ontologia *WordNet*, que melhor representa o conceito no domínio de interesse; (c) a extração de relações do significado mais representativo de cada conceito ; (d) a avaliação dos resultados obtidos. Os conceitos relevantes foram extraídos a partir de sentenças oriundas do corpus selecionado, com base em uma lista inicial de termos do domínio de nutrição. A partir da existência de termos do domínio de nutrição nos textos explicativos da *WordNet* foi escolhido o significado mais relevante de cada conceito. Os textos explicativos, associados aos significados selecionados, foram utilizados para a extração de relações. Foi realizada uma avaliação dos conceitos, dos significados e das relações resultantes do protótipo desenvolvido neste trabalho.

Palavras-chave: processamento de linguagem natural, automação de ontologias, extração de informação.

INTRODUÇÃO

O experimento utilizado neste trabalho é baseado em textos da língua inglesa. Por este motivo, todos os exemplos apresentados no decorrer deste texto, são extraídos de sentenças da língua inglesa.

A informação textual tem basicamente duas fontes. Fontes não estruturadas estão associadas a texto puro. Fontes semi-estruturadas seguem um determinado padrão, que facilita a automação da extração de informação dos textos (BUITELAAR; MAGNINI, 2005, p. 3-12). A Extração de Informação é uma área de pesquisa que trata da extração automática, a partir de fontes não estruturadas ou semi-estruturadas, de informação estruturada.

O estado da arte em Extração de Informações tem evoluído para a construção e extensão de ontologias (SARAWAGI, 2008, p. 261-377). Ontologias interligam entidades, relações e atributos em uma rede de informação para representar conhecimento de domínios gerais ou específicos (CARVALHEIRA, 2007, 142 p.). Entidades são conceitos de um dado domínio: *food, fruit, vitamin, disease*. Entidades podem ser associadas a seus atributos: *food – caloric value*. Relações de especialização são utilizadas para construir uma hierarquia de especialização, associando uma entidade mais específica a uma entidade mais genérica: *fruit – is – food*. Relações de associação vinculam entidades que não pertencem à mesma hierarquia de especialização: *vitamin – prevents – diseases*. Relações de agregação associam um todo a suas partes: *fruit – contains – seeds*.

A extração automática de ontologias gera uma base de conhecimento fundamental para a realização de pesquisas semânticas e extração de conhecimento a partir da imensa massa de informação textual da internet. Pesquisa Semântica vai além da utilização de palavras chaves e algoritmos baseados na estatística de uso de páginas na Internet, para a recuperação de textos cujo sentido tenha grande relação com questões formuladas por usuários (DONG; HUSSAIN; CHANG, 2008, p. 403-408).

Sistemas abertos (BANKO; ETZIONI, 2007, p. 95-102) tem sido propostos para suportar uma abordagem escalável para extração de relações em fontes não estruturadas da Web. *TextRunner* (BANKO at all, 2007, p. 2670-2676) foi o primeiro sistema a implementar a extração de várias relações em grandes corpus em um único passo de extração. Extratores anteriores realizavam a extração de um único tipo de relação alvo. Uma abordagem muito utilizada era o fornecimento de sementes de entrada, utilizadas, em geral, como pares de exemplos de uma relação alvo (BRIN, 1999, p. 172-183). O sistema *ReVerb* (FADER; SODERLAND; ETZIONI, 2011, p. 1535-1545) é uma evolução de *TextRunner*, para o qual é reportada uma maior acurácia na extração. O amadurecimento da pesquisa em extratores de informação é de extrema importância para viabilizar a complexa tarefa de construção e extensão automática de ontologias.

A *WordNet* (FELLBAUM, 1998, 422 p.) é uma ontologia, construída manualmente, que suporta informações sobre a semântica de substantivos, adjetivos e verbos da língua inglesa. Para cada substantivo, a WordNet provê, os vários sentidos associados ao substantivo e um texto explicativo (*gloss*) associado a cada sentido (*synset*) do substantivo em questão, bem como uma hierarquia de conceitos mais genéricos (*hypernyms*) e mais específicos (*hyponyms*).

A utilização da *WordNet*, para construção automática de ontologias é citada em diversos trabalhos na literatura. No sistema Alice (BANKO; ETZIONI, 2007, p. 95-102), que extende automaticamente uma ontologia construída manualmente, é utilizado um método de descoberta de conceito (entidade) que utiliza a *WordNet* para associar conceitos a uma dada categoria ou classe de entidade. *YAGO* (SUCHANEK; KASNECI; WEIKUM, 2008, p. 203-217) e *KOG Ontology* (WU; WELD, 2008, p. 635-644) são ontologias, construídas automaticamente a partir de fontes não estruturadas como a *Wikipedia*, que utililizam a *WordNet* para validar relações entre entidades.

Neste trabalho, foi investigado o processo de extração automática de novas relações a partir da *WordNet*, para posterior criação de uma ontologia núcleo. A extração é realizada a partir da utilização dos textos explicativos do sentido mais relevante para o domínio escolhido (nutrição), de um dado conceito (representado como um substantivo) da *WordNet*, resultando na extração de relações, de especialização e de associação, que serão utilizadas posteriormente para ampliar uma ontologia em construção (FAXINA; BATISTA, 2012, 9 p.) (ALMEIDA; BATISTA, 2013, 13 p.), resultante do projeto de pesquisa ao qual este trabalho está vinculado. O corpus utilizado é composto de sentenças extraídas automaticamente de um livro de referência para o ensino de nutrição básica (WHITNEY; ROLFES, 2011, 1007 p.).

1. DESENVOLVIMENTO

A metodologia utilizada incorpora quatro etapas: (a) a seleção automática de conceitos relevantes do domínio de nutrição; (b) a seleção automática do sentido, associado a cada conceito na ontologia WordNet, que melhor representa o conceito no domínio de interesse; (c) a extração de relações de especialização e de associação do sentido mais representativo de cada conceito; e (d) a avaliação dos resultados alcançados.

Neste artigo, a avaliação dos resultados alcançados será comentada ao final de cada seção, que descreve cada uma das três etapas do processo proposto. Os resultados apresentados foram extraídos de um protótipo, desenvolvido para validar o processo proposto. Com exceção das figuras geradas a partir do *WordNet Browser* (PRINCETON UNIVERSITY, 2014), as demais figuras foram extraídas de resultados gerados pelo protótipo desenvolvido neste trabalho.

1.1 SELEÇÃO DOS CONCEITOS RELEVANTES DO DOMÍNIO

A primeira atividade realizada foi a determinação de conceitos relevantes a partir do corpus. As sentenças, que foram utilizadas como corpus para este trabalho, foram extraídas (BATISTA; OLIVEIRA, 2014) dos quatro primeiros capítulos de um livro de referência na área de nutrição básica (WHITNEY; ROLFES, 2011, 1007 p.).

Em princípio, todos os substantivos existentes nas sentenças do corpus são candidatos a conceitos relevantes no domínio de nutrição. Das 1831 sentenças do corpus foram extraídos todos os substantivos, incluindo substantivos compostos como, por exemplo, "*balanced diet*". Para cada substantivo, extraído das sentenças do corpus, foi gerada a sua forma no singular e, então, verificada a sua existência na *WordNet*, resultando em um total de 1522 potenciais conceitos. Obviamente vários destes substantivos se repetem nas sentenças analisadas, e somente a forma singular de uma de suas ocorrências foi contabilizada como um potencial conceito distinto.

A maior dificuldade, neste passo da metodologia adotada, é a seleção automática dos conceitos relevantes para o domínio de nutrição. Alguns dos substantivos obtidos estão muito relacionados com o domínio de nutrição como, por exemplo: *"food"*, *"fruit"*, *"cereal"*, etc. Outros, no entanto, são muito genéricos, para serem caracterizados como relevantes para o domínio de nutrição como, por exemplo: *"week"*, *"worker"*, *"faculty member"*, etc.

A seleção manual seria um processo simples para um ser humano, e garantiria que os conceitos selecionados pertencem de fato ao domínio de interesse, mas descaracterizaria a construção de um processo automatizado. Além do mais a seleção manual, neste caso, implicaria na checagem em uma lista de 1522 substantivos. No caso de um corpus com uma ou mais ordens de grandeza maior, a seleção manual se tornaria inviável.

A solução adotada foi baseada na utilização de uma lista reduzida de palavras sementes para filtrar um número muito superior de conceitos potencialmente relevantes. Esta abordagem tem sido largamente utilizada na área de processamento de linguagem natural (BRIN, 1999, p. 172-183).

Para compor a lista de palavras sementes, foram adotados alguns substantivos intrinsecamente ligados ao domínio de nutrição e, adicionalmente, algumas de suas variantes. As palavras sementes são mostradas no quadro 1.

Adicionalmente, foram adotadas algumas palavras de descarte, por estarem relacionadas com termos comumente utilizados em livros ou por representarem conceitos um nível muito alto de abstração. As palavras escolhidas para descarte são mostradas no quadro 2.

Finalmente, foram selecionados, como conceitos relevantes para o domínio de nutrição, os substantivos que têm pelo menos a ocorrência de uma palavra semente no texto de algum de seus significados ou no texto explicativo de alguns de seus significados na *WordNet*, excluindo as palavras

de descarte. Como resultado, foram selecionados 332 conceitos, a partir da lista inicial de 1522 substantivos, para fazer parte do vocabulário do domínio de nutrição. A grande maioria dos conceitos selecionados é relevante para o domínio de nutrição, apesar de permanecerem alguns conceitos pouco significativos para o domínio como, por exemplo: "*sum*" e "*worker*". Adicionalmente, alguns substantivos relevantes como, por exemplo, "*lactose intolerance*" e "*alcohol abuse*" não foram selecionados para o vocabulário do domínio de nutrição.

QUADRO 1: Sementes para Seleção de Conceitos Relevantes no Domínio de Nutrição

nutrition – nutritious – nutrient
food
aliment – alimentation – alimentary
fruit
vegetable
meat
vitamin
protein
carbohydrate
diet
disease
body

QUADRO 2: Palavras de Descarte na Seleção de Conceitos Relevantes no Domínio de Nutrição

book – figure – text
substance – system

O maior problema na construção automática de ontologias é a incorporação de relações com semântica incorreta ou sem relação com o domínio de interesse. Baseado nesta premissa, a estratégia adotada é efetiva, porque a grande maioria dos conceitos selecionados está relacionada com o domínio de nutrição.

1.2 SELEÇÃO DO SIGNIFICADO MAIS RELEVANTE DE CADA CONCEITO

A segunda atividade realizada foi a seleção do significado, na *WordNet*, que melhor representa o conceito. Na figura 1, são ilustrados os vários significados atribuídos ao substantivo "*food*" na *WordNet*. No caso deste substantivo, o primeiro significado é o mais relacionado com o domínio de nutrição. Uma escolha ruim, do terceiro significado neste caso, resultaria na seleção de um texto explicativo sem nenhuma relação com o domínio de nutrição, comprometendo a extração de relações e, consequentemente, a construção automática da ontologia núcleo.

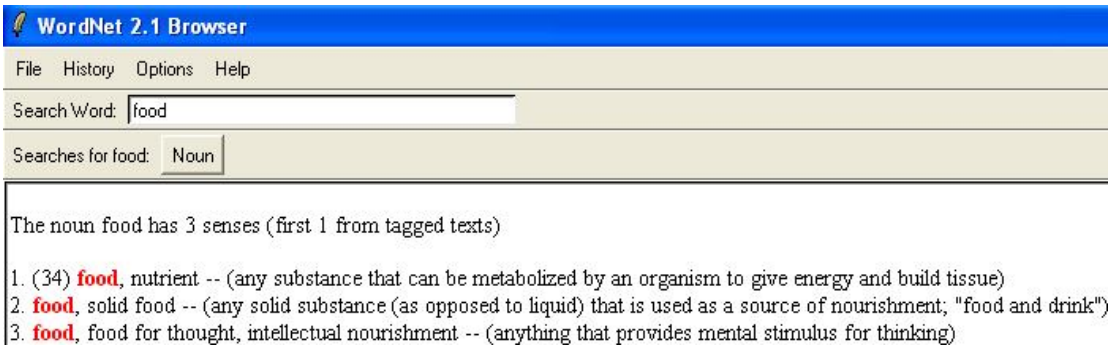

FIGURA 1: Os três significados do substantivo "*food*" na *WordNet* e seus textos explicativos

Existem alguns substantivos para os quais a *WordNet* representa um único significado como, por exemplo, o substantivo "*nutrient*". Na figura 2, é ilustrado o significado único deste substantivo, bem como, seu texto explicativo. No entanto, esta situação é minoritária, a maioria dos substantivos da *WordNet* tem mais de um significado, requerendo a seleção daquele que está mais relacionado com o domínio de interesse, para determinar quais os textos explicativos que serão utilizados para a extração de relações.

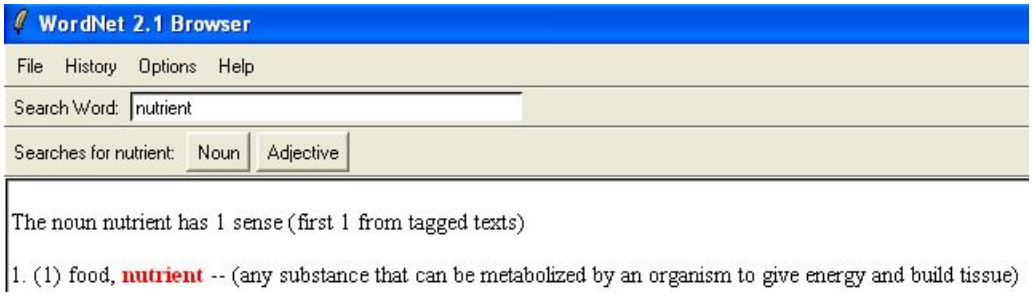

FIGURA 2: O único significado do substantivo "*nutrient*" na *WordNet* e seu texto explicativo

Foram testadas várias estratégias para a seleção do significado mais relevante de cada conceito no domínio de interesse. Neste trabalho, são apresentadas as duas estratégias que geraram os melhores resultados. A segunda estratégia reportada corresponde a uma adaptação da primeira estratégia.

A primeira estratégia se baseia nos seguintes critérios: (a) o maior total de substantivos no texto explicativo de um dado significado que pertencem ao vocabulário selecionado para o domínio; e (b) o maior total de sinônimos de um dado significado que pertencem ao vocabulário do domínio. O significado mais relevante para um dado conceito, do ponto de vista do domínio de interesse, é o significado que atende o primeiro critério e, em caso de empate, é o que atende ao segundo critério.

Na figura 3, são ilustrados os significados selecionados, bem como seus sinônimos e textos explicativos, para os cinco conceitos mais frequentes do vocabulário do domínio. A frequência, de cada conceito no vocabulário do domínio, foi calculada como sendo o número de ocorrências do substantivo, associado ao conceito, nas sentenças do corpus.

A primeira estratégia, para a seleção do significado mais relevante de cada conceito no domínio, gerou bons resultados na maioria dos casos. No entanto, ocorreram algumas situações nas

quais o conceito, mais relevante para o domínio de nutrição, não foi o selecionado. Na figura 4, são ilustrados os vários significados do substantivo "*fruit*" na *WordNet*. Para este conceito, o significado selecionado foi "*an amount of a product*", que não tem nenhuma relação com o domínio de nutrição, em detrimento da escolha correta, que deveria ter sido "*the ripened reproductive body of a seed plant*".

```
1: frequency 471
   food
      food - nutrient
         any substance that can be metabolized by an animal to give energy and build tissue
2: frequency 209
   nutrient
      food - nutrient
         any substance that can be metabolized by an animal to give energy and build tissue
3: frequency 159
   body
      body - organic structure - physical structure
         the entire structure of an organism (an animal, plant, or human being)
4: frequency 146
   sugar
      sugar - refined sugar
         a white crystalline carbohydrate used as a sweetener and preservative
5: frequency 140
   diet
      diet
         a prescribed selection of foods
```

FIGURA 3: Os significados selecionados para os cinco conceitos mais frequentes do vocabulário

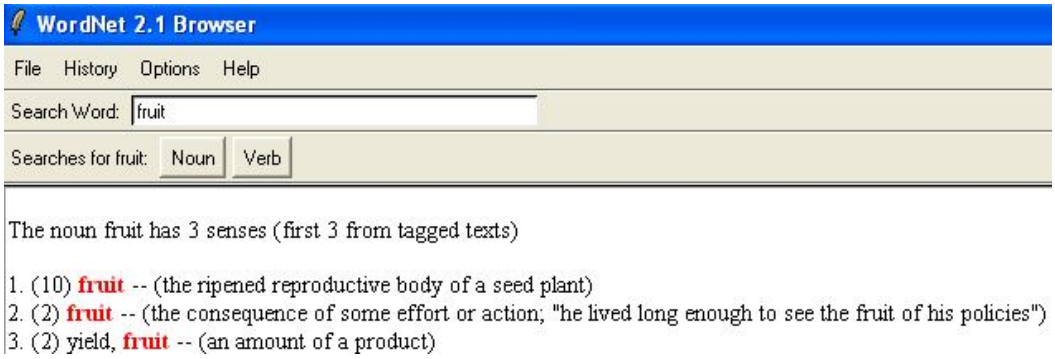

FIGURA 4: Os três significados do substantivo "*fruit*" na *WordNet* e seus textos explicativos

A segunda estratégia é uma adaptação da primeira. Na *WordNet*, os vários significados de um substantivo são informados em ordem decrescente em relação à frequência de ocorrência do respectivo substantivo na língua inglesa. A partir desta constatação, a segunda estratégia foi adaptada da seguinte forma: se o primeiro significado, do conceito do vocabulário, contiver pelo menos um conceito semente em seu texto explicativo, ele será selecionado como o significado mais relevante; caso contrário, será escolhido o significado que atender o primeiro critério da primeira estratégia.

Com a adoção da segunda estratégia, o significado selecionado para o conceito "*fruit*" foi o mais adequado para o domínio de nutrição, conforme ilustrado pela figura 5.

Apesar do fato da segunda estratégia apresentar melhores resultados, ainda permanecem alguns resultados inadequados como, por exemplo, para o conceito "*juice*", para o qual o significado selecionado mais relevante seria "*the liquid part that can be extracted from plant or animal tissue*", foi encontrado o resultado "*any of several liquids of the body*".

```
12: frequency 83
   milk
      milk
         a white nutritious liquid secreted by mammals and used as food by human beings
13: frequency 73
   intake
      consumption - ingestion - intake - uptake
         the process of taking food into the body through the mouth (as by eating)
14: frequency 71
   grain
      grain - caryopsis
         dry seed-like fruit produced by the cereal grasses: e.g. wheat, barley, Indian corn
15: frequency 70
   nutrition
      nutriment - nourishment - nutrition - sustenance - aliment - alimentation - victuals
         a source of materials to nourish the body
16: frequency 68
   fruit
      fruit
         the ripened reproductive body of a seed plant
17: frequency 66
   vitamin
      vitamin
         any of a group of organic substances essential in small quantities to normal metabolism
```

FIGURA 5: Os significados selecionados para seis conceitos, incluindo o conceito "*fruit*".

1.3 EXTRAÇÃO DE RELAÇÕES DOS TEXTOS EXPLICATIVOS

A terceira atividade realizada foi a extração das relações de cada uma das partes do texto explicativo referente ao significado selecionado como mais relevante para cada um dos conceitos selecionados para o domínio de nutrição.

A representação dos textos explicativos apresenta várias variações, e para cada uma delas é necessário montar uma composição distinta de sentença, a partir da qual foram extraídas relações de especialização e de associação. Na figura 6, são ilustrados conceitos, cujos textos explicativos exemplificam as variações dos textos, encontrados na *WordNet*, que precisam ser considerados para formar as sentenças das quais serão extraídas as relações.

Com base nos textos explicativos, ilustrados na figura 6, é possível determinar quatro situações distintas, que precisam ser consideradas para a composição de sentenças. Para textos explicativos iniciados com substantivos, a sentença é composta por: <conceito> "*is*" <texto explicativo>. Para textos explicativos iniciados com um pronome possessivo, a sentença é composta por: <conceito> <texto explicativo sem o pronome possessivo inicial>. Para textos explicativos iniciados com verbos, a sentença é composta por: <conceito> <texto explicativo>. Para explicativos iniciados com a área de interesse, para a descrição do conceito, entre parênteses, a sentença é composta por: <conceito> <texto explicativo sem a área de interesse>. No quadro 3, são ilustradas as sentenças geradas, a partir dos conceitos ilustrados na figura 6, para as quatro situações de composição das sentenças.

A partir dos textos explicativos, encontrados na *WordNet*, para descrever os significados selecionados dos 332 conceitos do vocabulário, foram geradas 401 sentenças, para servir de entrada para a extração de relações. O número de sentenças geradas é maior do que o número de conceitos,

porque alguns textos explicativos são compostos de vários trechos de texto, como exemplificado na figura 6 para o substantivo "*carbohydrate*".

```
noun: fiber
   roughage - fiber
       coarse, indigestible plant food low in nutrients
       its bulk stimulates intestinal peristalsis
noun: curry
       (East Indian cookery) a pungent dish of vegetables or meats flavored with curry powder
           and usually eaten with rice
noun: immunity
   immunity - resistance
       (medicine) the condition in which an organism can resist disease
noun: carbohydrate
   carbohydrate - saccharide - sugar
       an essential structural component of living cells and source of energy for animals
       includes simple sugars with small molecules as well as macromolecular substances
       are classified according to the number of monosaccharide groups they contain
```

FIGURA 6: Conceitos que ilustram variantes de textos explicativos da *WordNet*

QUADRO 3: Sentenças compostas a partir dos textos explicativos do conceitos da figura 4

Texto explicativo iniciado com substantivo
fiber is coarse, indigestible plant food low in nutrients.
carbohydrate is an essential structural component of living cells and source of energy for
Texto explicativo iniciado com pronome possessivo (ex: its)
fiber bulk stimulates intestinal peristalsis
Texto explicativo iniciado com verbo
carbohydrate includes simple sugars with small molecules as well as macromolecular
carbohydrates are classified according to the number of monosaccharide groups they
Texto explicativo iniciado com a área de interesse entre parênteses
curry is a pungent dish of vegetables or meats flavored with curry powder and usually
immunity is the condition in which an organism can resist disease.

Foi implementado um extrator para separar cada uma das sentenças em relações, cujas principais regras de extração são baseadas nos seguintes separadores: "*that*" atuando como pronome relativo; "*and*" e "*or*" atuando como conjunção coordenada; e verbos no particípio passado após substantivos . Na figura 7, são ilustradas as relações, extraídas da sentença criada a partir de um texto explicativo do conceito "food", geradas a partir do tratamento dos separadores "*that*" e "*and*".

```
noun: food
   food is any substance that can be metabolized by an animal to give energy and build tissue.
      food - is - substance
      food - can be metabolized by - animal
      food - give - energy
      food - build - tissue
```

FIGURA 7: Relações extraídas a partir do texto explicativo do conceito "*food*"

Na figura 8, são ilustradas as relações, extraídas da sentença criada a partir de um texto explicativo do conceito "*sugar*" para ilustrar as extrações, geradas a partir do tratamento dos separadores: verbo no particípio passado após substantivo "*carbohydrate used*" e "*and*".

Em todas as relações, ilustradas nas figuras 7, 8 e 9, a determinação da primeira entidade da relação binária é simples, pois corresponde ao substantivo que antecede o verbo da sentença da relação. Em geral, a primeira entidade da relação é o substantivo correspondente ao próprio conceito que foi

utilizado para compor a sentença do texto explicativo, a partir da qual a relação foi extraída. Em alguns casos, este substantivo pode estar adjetivado como, por exemplo, "*fiber bulk*", gerado para substituir o pronome possessivo "*its*", conforme ilustrado no quadro 3.

noun: sugar
sugar is a white crystalline carbohydrate used as a sweetener and preservative.
- sugar - is - white crystalline carbohydrate
- sugar - is used as - sweetener
- sugar - is used as - preservative

FIGURA 8: Relações extraídas a partir do texto explicativo do conceito "*sugar*"

noun: carbohydrate
- carbohydrate is an essential structural component of living cells and source of energy for animals.
--- carbohydrate - is - essential structural component of living cells
--- carbohydrate - is - source of energy for animals
- carbohydrate includes simple sugars with small molecules as well as macromolecular substances.
--- carbohydrate - includes - simple sugars with small molecules
--- carbohydrate - includes - simple sugars with macromolecular substances
- carbohydrate are classified according to the number of monosaccharide groups they contain.
--- carbohydrate - are classified according to - number of monosaccharide groups
--- carbohydrate - contain - monosaccharide groups

FIGURA 9: Relações extraídas a partir do texto explicativo do conceito "*sugar*"

No entanto, a determinação da segunda entidade da relação não é simples, no seu caso mais geral. Nas relações ilustradas nas figuras 7 e 8, a caracterização da segunda entidade da relação é simplesmente o substantivo adjetivado que sucede o verbo da sentença. No entanto, nas relações ilustradas na figura 9, a frase verbal da relação é sucedida por uma frase substantiva composta por um encadeamento de substantivos e preposições. Nestes casos, a tarefa de determinar a segunda entidade da relação é complexa. Para a frase substantiva, candidata à segunda entidade, "*essential structural component of living cells*", a escolha do substantivo adjetivado, que inicia o encadeamento, parece ser a escolha mais adequada, embora resulte na perda da informação que especifica a entidade selecionada: "*of living cells*". Considerando, no entanto, a frase substantiva "*source of energy for animals*", a escolha somente do primeiro substantivo do encadeamento, ou seja "*source*", resulta em uma grande perda de semântica para a relação. No caso das frases substantivas "*simple sugars with small moleculas*" e "*simple sugar with macromolecular substances*", a escolha dos primeiro substantivos adjetivados, de cada encadeamento, resulta em uma segunda entidade idêntica para as duas relações. A determinação da segunda entidade de uma relação, a partir de frase substantivas representadas por um encadeamento de substantivos e preposições, é uma tarefa complexa que está fora do escopo deste trabalho.

CONSIDERAÇÕES FINAIS

Neste trabalho foi desenvolvido um protótipo, a partir do processo proposto, que suporta a extração de relações a partir de textos explicativos da *WordNet*, associados aos significados relevantes

de conceitos extraídos de sentenças em linguagem natural. A proposta central deste trabalho é a utilização de textos explicativos da *WordNet* para extrair relações no domínio de nutrição, para posterior utilização na construção automática de ontologias. Os textos explicativos da *WordNet* foram gerados manualmente e fornecem uma fonte muito mais confiável para a extração de relações, do que sentenças extraídas de textos da internet.

A grande vantagem do processo, proposto neste trabalho, é que suas três etapas foram automatizadas. No entanto, os resultados gerados, em cada uma de suas etapas, podem ser aperfeiçoados em trabalhos futuros.

No vocabulário do domínio, gerado como resultado da primeira etapa deste processo, são incluídos alguns conceitos muito gerais para o domínio de nutrição, embora a grande maioria dos conceitos selecionados seja relevante Sob outro ponto de vista, alguns substantivos, extraídos das sentenças do corpus, deveriam ter sido identificados como conceitos relevantes do domínio de interesse. Um trabalho futuro, para aperfeiçoar este etapa do processo seria complementar a utilização de conceitos sementes, com outras estratégias para a seleção dos conceitos relevantes do domínio de interesse.

Na seleção do significado mais relevante para cada conceito na *WordNet*, realizada na segunda etapa do processo, ocorrem alguns poucos casos no qual o significado mais adequado não é o selecionado. Estas exceções indicam também a necessidade de aperfeiçoar a estratégia proposta para a seleção do significado mais relevante para cada conceito.

Finalmente, para as relações, geradas na terceira etapa deste processo, será necessário pesquisar uma etapa adicional, para extrair a segunda entidade das relações, nas situações que envolvem frases substantivas compostas por encadeamentos de substantivos e preposições.

REFERÊNCIAS

ALMEIDA, Francy Helder Silva; BATISTA, Joinvile Junior. Protótipo para construção de ontologia núcleo a partir de fonte não estruturada. ENEPE, 2013, 13 p.

BANKO, Michele; ETZIONI, Oren. Strategies for lifelong knowledge extraction from the Web. In: Proceedings of the 4th international conference on Knowledge Capture, 2007, p. 95-102.

BANKO, Michele; CAFARELLA. Michael; SODERLAND, Stephen; BROADHEAD, Matt; ETZIONI, Oren. Open information extraction from the web. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2007, p. 2670-2676.

BATISTA, Joinvile Junior; OLIVEIRA, Paulo Edson Misokami. Protótipo para construção de ontologia núcleo a partir de fonte não estruturada. ENEPE, 2014.

BRIN, Sergei. Extracting patterns and relations from the World Wide Web. In Selected papers from the International Workshop on the WWW and Databases, 1999, p. 172-183.

BUITELAAR, Paul; MAGNINI, Bernardo. Ontology learning from text: An overview. Ontology learning from text: Methods, evaluation and applications, v. 123 of Frontiers in Artificial Intelligence and Applications. IOS Press, 2005, p. 3-12.

CARVALHEIRA, Luiz Carlos Cruz. Método semi-automático de construção de ontologias parciais de domínio com base em textos. Dissertação de Mestrado, Escola Politécnica de São Paulo, 2007, 142 p.

DONG, Hai; HUSSAIN, Farookh Khadeer; CHANG, Elizabeth. A Survey in Semantic Search Technologies. In: Second IEEE International Conference On Digital Ecosystems and Technologies, 2008, p. 403-408.

FADER, Anthony; SODERLAND, Stephen; ETZIONI, Oren. Identifying Relations for Open Information Extraction. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011, p. 1535-1545.

FAXINA, Giuliano; BATISTA, Joinvile Junior. Protótipo para construção de ontologia núcleo a partir de fonte não estruturada. ENEPE, 2012, 9 p.

FELLBAUM, Christiane. WordNet: An Electronic Lexical Database. MIT Press, 1998, 422 p.

PRINCETON UNIVERSITY, WordNet Browser 2.1. WordNet – A lexical database for English. http://wordnet.princeton.edu/wordnet/download/current-version/. Acesso em 2014.

SARAWAGI, Sunita. Information Extraction. Foundations and Trends in Databases: Vol. 1: No 3, 2008, p. 261-377.

SUCHANEK, Fabian M.; KASNECI, Gjergji; WEIKUM, Gerhard. YAGO: A Large Ontology from Wikipedia and WordNet. Elsevier Journal of Websemantics, 2008, p. 203-217.

WHITNEY Ellie; ROLFES Sharon Rady. Understanding Nutrition; Twelfth Edition; 2011, 1007 p. Wadsworth Cengage Learning.

WU, Fei; WELD, Daniel S. Automatically refining the Wikipedia infobox ontology. In Proceedings of the International Conference on World Wide Web, 2008, p. 635-644.