Alinhamento de Grafos: Investigação do Alinhamento de ConceptNets para a Tradução Automática

Paulo Henrique Barchi[1], **Helena de Medeiros Caseli**[1]
Junia Coutinho Anacleto[1]

[1]Departamento de Computação, Universidade Federal de São Carlos (UFSCar)
Rod. Washington Luís, Km 235, CP 676, CEP 13565-905 São Carlos-SP

{paulobarchi@comp,helenacaseli@dc,junia@dc}.ufscar.br

Abstract. *This paper describes a research proposal to align concepts in parallel semantic networks, particularly in Brazilian Portuguese and English ones. The semantic networks (ConceptNets) under study in this proposal are composed of nodes and arcs (to connect nodes). The nodes represent the knowledge derived from the common sense base while the arcs represent relations between two nodes based on studies on the theory of [Minsky 1986]. After the alignment, the aligned similar concepts in different languages could be used to extract useful knowledge for machine translation.*

Resumo. *Esse artigo descreve uma proposta de pesquisa que visa o alinhamento de conceitos em redes semânticas paralelas, particularmente para os idiomas português do Brasil e inglês. As redes semânticas (ConceptNets) consideradas nesta proposta estão estruturadas em nós e arcos (que conectam os nós). Os nós armazenam os conhecimentos da base de senso comum, enquanto os arcos representam as relações entre dois nós, baseadas nos estudos sobre a teoria de [Minsky 1986]. A partir desse alinhamento de conceitos similares em idiomas distintos outras técnicas poderão ser aplicadas para extração de conhecimento útil para a tradução automática.*

1. Introdução

A Tradução Automática (TA), uma das áreas de pesquisa mais antigas e mais fortes de Inteligência Artificial, pode ser entendida como a tradução de uma língua natural (fonte) para outra (alvo) por meio de programas de computador. Um dos grandes obstáculos dessa área é gerar um texto-alvo com o significado mais próximo possível daquele originalmente existente no texto-fonte.

Nesse contexto, o trabalho aqui apresentado está inserido num projeto maior que visa a investigação do uso de senso comum na tradução automática, inicialmente, do par de idiomas português do Brasil e inglês. Para tanto, o trabalho aqui apresentado tem como objetivo a realização de um primeiro passo fundamental para o uso do conhecimento representado por duas redes semânticas (ConceptNets) de idiomas distintos: o alinhamento dessas redes, ou seja, a identificação dos mapeamentos entre os nós em uma rede com os nós da outra rede.

Assim, para contextualizar essa pesquisa, a seção 2 apresenta um método de alinhamento entre duas redes semânticas paralelas que poderá servir de base para esse trabalho. Na seção 3, descreve-se, brevemente, a estrutura das ConceptNets que este trabalho pretende alinhar. Por fim, a seção 4 conclui este documento.

2. Levantamento Bibliográfico

Em [Carpuat et al. 2006] os autores propõem uma aproximação para o alinhamento de ontologias no nível de nós: dado um conceito representado por uma palavra de sentido particular em uma ontologia, a tarefa é encontrar a melhor correspondência de palavra em uma segunda ontologia. Para tanto, os autores apresentam um novo método independente de língua, baseado em corpus, que utiliza técnicas de recuperação de informação e tradução automática.

O interesse dos autores de tal artigo é mapear nós que descrevem conceitos similares, em ontologias em diferentes línguas, um objetivo similar ao da proposta aqui apresentada na qual os nós estão em ConceptNets em diferentes línguas. Ao contrário de alinhar palavras únicas ou frases, a ideia é mapear grupos de palavras ou frases que são sinônimos em uma língua com um grupo de palavras na outra língua que são empregadas no mesmo sentido.

Para comparar os nós das duas ontologias estudadas em [Carpuat et al. 2006], WordNet e HowNet, adotou-se uma representação comum simples na qual define-se um nó como um conjunto de palavras que compartilham uma definição comum. Assim, um nó N pode ser representado por dois conjuntos de palavras: N = (S, D), em que S representa o conjunto de *sinônimos* da palavra de sentido N e D é o conjunto de palavras de *definições* para N.

No caso de [Carpuat et al. 2006], o mapeamento dos nós é realizado com base na tradução extraída de um dicionário bilíngue e também em um algoritmo que mapeia os nós de uma rede, candidatos ao alinhamento final, com cada nó correspondente na outra rede. Nesse algoritmo, dado um nó da HowNet, $N_{HowNet} = (S_{HowNet}, D_{HowNet})$, deve-se definir: (i) um conjunto contendo as traduções das palavras em S_{HowNet}: $E = \{e_1, e_2, \ldots, e_m\}$ tal que e_i é a tradução de pelo menos uma palavra em S_{HowNet} e (ii) o conjunto de candidatos ao alinhamento $C = \{N_{WordNet,1}, \ldots, N_{WordNet,p}\}$ tal que para cada nó da WordNet $N_{WordNet,i} = (S_{WordNet,i}, D_{WordNet,i})$, existe ao menos um e_j que pertence a $S_{WordNet,i}$.

O mapeamento inicial conta com uma lista de traduções que definem relacionamentos entre palavras. Para caracterizar cada nó mais precisamente e para ter mais informações para diferenciá-lo de outros nós, é necessário definir características que coletam informações contextuais. Para isso, os autores propõem um método (baseado em corpus) inspirado pelo algoritmo de Convec [Fung e Lo 1998] apud [Carpuat et al. 2006]. A similaridade métrica entre palavras em duas línguas diferentes é determinada capturando-se a co-ocorrência das palavras em um corpus com uma lista de *seed-words* para as quais a tradução em ambas as línguas está definida. A similaridade de um par de palavras ($sim(w, v)$) é definida com base no cosseno dos vetores de contexto dessas palavras.

Cada elemento do vetor é um peso que corresponde a uma função de significância de uma *seed-word* particular e sua frequência de co-ocorrência com a palavra de interesse. Dado o par de línguas, primeiro define-se um conjunto SW de pares de *seed-words*: $SW = (s_{10}, s_{20}), (s_{11}, s_{21}), \ldots, (s_{1t}, s_{2t})$, tal que cada palavra s_{2i} é uma tradução de s_{1i}. O conjunto de pares de *seed-word* SW é definido uma única vez e para todas as palavras para um par de línguas. Por fim, a similaridade entre um nó da WordNet ($N_{WordNet} = (S_{WordNet}, D_{WordNet})$) e um da HowNet ($N_{HowNet} = (S_{HowNet}, D_{HowNet})$)

é definida como:

$$sim(N_{WordNet}, N_{HowNet}) = \frac{(sim(S_{WordNet}, S_{HowNet}) + sim(D_{WordNet}, D_{HowNet}))}{2}$$

Eventualmente os nós candidatos de C são ordenados de acordo com a similaridade com o nó N_{HowNet}, e o synset com a maior similaridade é considerado o "vencedor" do alinhamento.

Como visto, a proposta de [Carpuat et al. 2006] é alinhar os nós de duas ontologias em línguas diferentes, usando informações de tradução e similaridade calculada estatisticamente. A avaliação de tal método é obtida julgando-se os resultados como corretos ou incorretos, verificando-se se a tradução dada é realmente a esperada, o que pode ser feito por juízes humanos familiarizados com as línguas em questão.

3. Alinhamento em Redes de Conceitos (ConceptNets)

As redes semânticas (ConceptNets) consideradas nesta proposta diferem das ontologias de [Carpuat et al. 2006], mas apresentam estrutura similar de nós e arcos (que conectam os nós). Os nós armazenam os conhecimentos da base de senso comum, enquanto os arcos representam as relações entre dois nós. As relações foram definidas baseadas nos estudos sobre a teoria de [Minsky 1986] de como o pensamento humano funciona.

Existem 17 relações da ConceptNet brasileira [Anacleto et al. 2008] compartilhadas com a norteamericana [Singh 2002], como CapableOf, UsedFor, IsA, PartOf, DefinedAs e MadeOf. A figura 1 ilustra um trecho da ConceptNet norteamericana.[1]

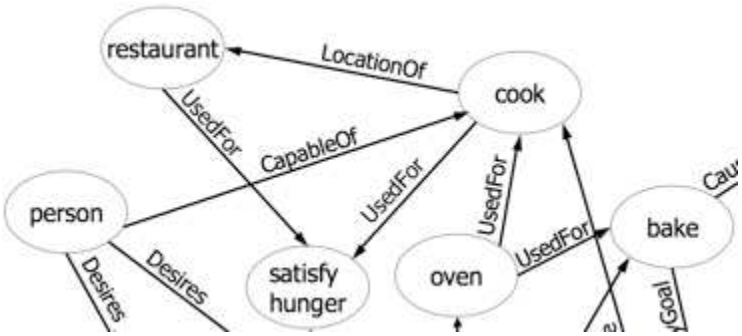

Figura 1. Trecho da ConceptNet norteamericana

O trabalho aqui apresentado tem como objetivo a realização do alinhamento dessas redes, ou seja, a identificação dos mapeamentos entre os nós em uma ConceptNet com os nós de outra ConceptNet, ambas com estrutura similar à apresentada na figura 1. Para tanto, pretende-se explorar as correspondências entre as palavras que representam os conceitos em cada nó (obtidas de um dicionário bilíngue estatístico compilado automaticamente) e também a estrutura hierárquica dos conceitos, além da métrica de similaridade considerada no método de [Carpuat et al. 2006].

Para ilustrar essa ideia, por exemplo, considere o trecho da ConceptNet apresentado na figura 1 e as possíveis traduções da palavra para o inglês *oven* disponíveis no dicionário bilíngue estatístico inglês-português compilado automaticamente com base em corpus: *forno*, *estufa*, *frio* e *fogo*, ordenadas da mais provável para a menos provável. A

[1]Trecho extraído da figura disponível em `http://conceptnet.media.mit.edu/`.

partir da correspondência entre a palavra *oven* e a palavra *forno* seria possível estabelecer um alinhamento do trecho da ConceptNet em inglês apresentado na figura 1 e um trecho equivalente contendo a palavra *forno* na ConceptNet em português.

Um processo semelhante poderia levar ao alinhamento das palavras *cook* em inglês e *cozinhar* em português. Desses alinhamentos poderia-se derivar conhecimento do tipo "um fogão é usado para cozinhar", ou seja, mesmo que essa informação não esteja explícita na ConceptNet em português, é possível obtê-la a partir dos alinhamentos citados.

Esse é apenas um exemplo de como o alinhamento de palavras poderia ser diretamente mapeado no alinhamento de conceitos nas ConceptNets. Nos casos em que tal mapeamento não seja possível, outras informações como a estrutura hierárquica dos conceitos, poderão ser usadas para a realização do alinhamento.

4. Considerações Finais

Com o desenvolvimento de uma ferramenta de alinhamento de ConceptNets paralelas, ou mais genericamente de estruturas de grafos paralelas, espera-se fornecer à área de tradução automática e outras que trabalham com processamento multilíngue (por exemplo, a recuperação de informação multilíngue e *cross-language*) um novo recurso para processamento de exemplos multilíngues.

Uma vez que não há conhecimento de métodos de alinhamento das ConceptNets citadas ou mesmo estruturas de grafos nos moldes estudados nesse trabalho, acredita-se que os resultados dessa pesquisa serão muito importantes e mesmo inovadores para diversas áreas que trabalham com grafos, resultando em uma variedade de aplicações. A implementação de tal método de alinhamento é o próximo passo nesse projeto de Iniciação Científica em fase inicial de desenvolvimento.

Referências

J. C. Anacleto, A. F. P. Carvalho, E. N. Pereira, A. M. Ferreira e A. J. F. Carlos (2008). Machines with good sense: How can computers become capable of sensible reasoning? Em *IFIP AI*, páginas 195–204.

Marine Carpuat, Pascale Fung e Grace Ngai (2006). Aligning word senses using bilingual corpora. *ACM Transactions on Asian Language Information Processing (TALIP)*, 5(2):89–120.

P. Fung e Y. Y. Lo (1998). An ir approach for translating new words from nonparallel, comparable texts. Em *Proceedings of the 36th Annual Conference of the Association for Computational Linguistics*, páginas 414–420, Montreal, Canada.

M. Minsky (1986). *The Society of Mind*. Simon and Schuster, New York.

P. Singh (2002). The OpenMind Commonsense project. KurzweilAI.net. Disponível em: <http://web.media.mit.edu/ push/OMCSProject.pdf>.