Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

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

Paulo Henrique Barchi
Helena de Medeiros Caseli

NILC-TR-10-07

Agosto, 2010

Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

Esse relatório descreve uma pesquisa desenvolvida com o intuito de alinhar os conceitos em redes semânticas paralelas, particularmente para os idiomas português do Brasil e inglês. As redes semânticas (ConceptNets) consideradas nesse trabalho armazenam o conhecimento de senso comum e estão estruturadas em nós e arcos (que conectamos nós por meio de "relações semânticas"). O senso comum pode ser definido como "o conhecimento compartilhado por um determinado grupo de pessoas em um dado tempo, espaço e cultura" (SINGH, 2002). Essa fonte rica de informações pode ser representada por redes semânticas (ConceptNets) e pode ser muito útil para muitas aplicações da área de Processamento de Linguagem Natural (PLN) como: a recuperação de informação (RI) multilíngue e *cross-language* e a tradução automática (TA). A tradução automática, por sua vez, pode ser entendida como a tradução de uma língua natural (fonte) para outra (alvo), tendo como um dos grandes desafios a geração de um texto-alvo que preserve o significado presente no texto-fonte (NIRENBURG, 1987). Nesse contexto, esse relatório descreve a investigação de um primeiro passo rumo ao uso de senso comum na TA: o alinhamento entre duas ConceptNets (americana e brasileira), ou seja, o mapeamento dos conceitos representados em uma rede com os conceitos representados na outra rede. Para a realização dessa tarefa, foram desenvolvidos 5 modos de alinhamento (A, B1, B2, B3 e C) que realizam o alinhamento basicamente considerando (i) as traduções presentes em um léxico bilíngue gerado automaticamente a partir de *corpus* paralelo e (ii) a estrutura das ConceptNets, ou seja, o número e o tipo das relações existentes entre os nós. A análise manual de uma amostra dos alinhamentos gerados por cada modo mostrou que os melhores critérios de alinhamento são: as traduções presentes no léxico e a equivalência entre as relações que envolvem os nós. Dessa maneira, com o desenvolvimento dessa ferramenta de alinhamento de ConceptNets paralelas (ou estruturas de grafos paralelas), espera-se fornecer à área de TA e outras que trabalham com processamento multilíngue (RI multilíngue e *cross-language*) um novo recurso.

Índice

Alinhamento de Grafos: Investigação do Alinhamento de ConceptNets para a Tradução Automática[1]

1 Introdução

Uma das grandes dificuldades das aplicações multilíngues como a Tradução Automática ou a Recuperação de Informação multilíngue é o mapeamento, de uma língua para outra, do significado atribuído a um conceito. Para tanto, é necessário o processamento de conhecimento semântico e/ou contextual, o qual pode ser representado de várias maneiras. Nesse trabalho, considera-se que os conceitos em uma língua estão representados em redes (grafos) semânticas de conceitos denominadas ConceptNets, nas quais os conceitos são os nós e as relações entre eles, os arcos que os conectam. Desse modo, cada nó contém uma palavra ou um conjunto de palavras e uma relação entre dois nós representa o relacionamento existente entre eles definido de acordo com a teoria de Minsky (1986). Em uma ConceptNet é possível representar, por exemplo, que os conceitos "livro" e "aprender" estão relacionados por meio da relação UsedFor (usado para) indicando que "livro" é usado para "aprender", o que é representado como UsedFor(livro,aprender).

O conhecimento armazenado nas ConceptNets é considerado conhecimento de Senso Comum, o que pode ser definido como "o conhecimento compartilhado por um determinado grupo de pessoas em um dado tempo, espaço e cultura" (SINGH, 2002). Nesse sentido, o trabalho aqui descrito se baseia na hipótese de que um recurso que relacione os conceitos de uma língua com os conceitos de outra língua seria de extrema relevância para auxiliar aplicações multilíngue como as citadas. Como um passo inicial para verificar tal hipótese, esse trabalho apresenta experimentos realizados no intuito de determinar o alinhamento (mapeamento) de conceitos em duas ConceptNets paralelas. As ConceptNets utilizadas nos experimentos aqui apresentados foram geradas a partir de bases de senso comum propostas e adotadas nos projetos *Open Mind Common Sense* OMCS[2] (SINGH, 2002) e OMCS-Br[3] (ANACLETO et al., 2006, 2008) contendo conceitos em inglês e português do Brasil, respectivamente.

A proposta, a longo prazo, é determinar como e em que medida a informação de senso comum, representada em duas ConceptNets para idiomas diferentes, pode auxiliar o processo

[1] Este trabalho foi apoiado por CNPq/PIBIC/UFSCar e FAPESP.
[2] http://commons.media.mit.edu/en/
[3] http://www.sensocomum.ufscar.br

de tradução automática envolvendo tais idiomas. A Tradução Automática (TA) é uma das áreas de pesquisa mais antigas e mais fortes de Inteligência Artificial e pode ser entendida como a tradução de uma língua natural (fonte) para outra (alvo) por meio de programas de computador. A tarefa de TA consiste em, dado um texto escrito em uma língua de origem (texto-fonte), produzir uma versão do mesmo texto em outra língua natural (texto-alvo). O grande problema desse processo, segundo (NIRENBURG, 1987), é gerar, como saída, uma versão (texto-alvo) que mantenha o significado mais próximo possível daquele originalmente existente no texto-fonte. Para alguns autores, inclusive, a tradução humana é considerada uma arte (HUTCHINS, 1998) já que envolve escolhas pessoais, não sendo simplesmente uma questão de substituições diretas de sequências de símbolos (DIAS DA SILVA et al., 2007). É exatamente com o intuito de auxiliar no mapeamento de significados no processo de tradução que esse trabalho foi desenvolvido.

Assim, esse trabalho descreve um primeiro passo fundamental para o uso, na TA, do conhecimento representado por duas ConceptNets de idiomas distintos: o alinhamento dessas redes, ou seja, a identificação das correspondências entre os conceitos (nós) em uma rede e os conceitos (nós) da outra rede. Nos experimentos aqui apresentados, tal mapeamento é realizado entre conceitos em português do Brasil e inglês, porém a abordagem proposta é independente de língua sendo dependente apenas da estrutura das ConceptNets, o que pode ser facilmente adaptado a outras redes ou grafos representando conhecimentos paralelos.

O levantamento bibliográfico realizado para embasar essa pesquisa e dar origem aos modos de alinhamento de ConceptNets propostos é apresentado na seção 2, enquanto que a seção 3 descreve os recursos (ConceptNets e léxico bilíngue) usados nos experimentos aqui apresentados. A seção 4 traz a explicação de cada um dos 5 modos de alinhamento propostos, seguida (na seção 5) dos resultados de experimentos para a avaliação dos mesmos. Por fim, a seção 6 apresenta algumas considerações finais a respeito desse trabalho.

2 Levantamento Bibliográfico

Para a realização desse trabalho pesquisou-se métodos de alinhamento existentes para alinhar estruturas de grafos ou similares (árvores, por exemplo) já que nenhum trabalho específico para alinhamento de ConceptNets foi encontrado na literatura.

Em (TINSLEY et al., 2009), por exemplo, os autores descrevem como é feito o uso de novas ferramentas para a construção automática de uma grande árvore sintática e extração de um grupo de pares de frase linguisticamente determinadas da árvore. Os mesmos autores apresentam em (TINSLEY et al., 2007) um algoritmo que induz alinhamentos entre estruturas pareadas linguisticamente dos quais uma ordem de formas superficiais de constituintes pode ser determinada. Em (ZHECHEV, 2009) é introduzido um sistema open-source para uma geração automática rápida e robusta de *treebanks* paralelas com ênfase ao alinhamento subárvore. Outros trabalhos também tratam do alinhamento ou construção de conjunto de árvores sintáticas alinhadas ou não como em (SAMUELSSON et al., 2008), no qual os autores relatam a construção de uma *treebank* paralela trilíngue em inglês, alemão e sueco com aproximadamente 500 árvores do romance “O mundo de Sofia” e outras 500 de textos de economia. De dois dos três autores de (SAMUELSSON et al., 2008), o artigo de Samuelsson e Volk (2007) aborda a mesma área de pesquisa (*treebank* paralela). Ainda nessa área, em (LUNDBORG et al., 2007), são apresentados vários casos de uso para o Stockholm TreeAligner, uma ferramenta de software originalmente feita para anotar os alinhamentos em uma *treebank* paralela.

Embora tratem do alinhamento e manipulação de estruturas de grafos (nesse caso árvores sintáticas) diferentes da investigada nesse trabalho (redes semânticas de conceitos), o conhecimento adquirido com essas leituras foi bastante interessante e ajudou a delimitar o que se faz em áreas correlatas. O mesmo ocorre com o trabalho (DOERR & IORIZZO, 2008), no qual os autores frisam a dificuldade de se fazer o mapeamento entre ontologias, que vem sendo pesquisado há décadas, e oferecem uma nova aproximação para tal realização. No entanto, não apresentação nenhuma implementação propriamente dita. De modo semelhante, em (NAVIGLI, 2009) é apresentada uma proposta para desambiguação das palavras de um dado texto a partir do mapeamento dessas com seus sentidos do dicionário, mas não apresenta nenhum algoritmo.

Outras estruturas de conceitos também foram investigadas como a EuroWordNet apresentada em (PETERS et al., 1998) na qual bases de dados semânticas como a WordNet1.5 para várias línguas são combinados por uma chamada inter-index-lingual, relacionada ao alinhamento de redes semânticas distintas das investigadas nesse trabalho. A nova versão da ConceptNet (3.0) é apresentada por Havasi et al. (2009) juntamente com o projeto *Open Mind Common Sense*. Essa nova versão melhora a aquisição de novos conhecimentos com novas relações.

Finalmente, fortemente relacionado com o propósito desse trabalho, em (CHUNG et al., 2007), é apresentado o projeto GlobalMind que visa computar automaticamente diferenças culturais e similaridades para a tradução automática. GlobalMind fornece banco de dados de senso comum (similar à ConceptNet) de vários países e linguagens além de dois módulos de inferência para analisar e computar diferenças culturais e similaridades do banco de dados. A partir da descrição do módulo de inferência de conceito similar é que surgiu a ideia inicial para o alinhamento das ConceptNets, descrita abaixo.

Pelo estudo dos artigos acima, foi possível elaborar uma ideia inicial de como realizar o alinhamento das ConceptNets. Inicialmente, pensou-se em utilizar os alinhamentos de palavras gerados por ferramentas automáticas estatísticas como NATools[4] e GIZA++[5] (toolkits muito utilizados para tradução automática estatística).

Dado o alinhamento inicial de palavras, o nó alinhado é considerado a semente para a expansão realizada de modo similar ao SIM (*Similar Inference Module* - Módulo de inferência de conceito similar) proposto em (CHUNG et al., 2007) que tem como objetivo encontrar os dois nós (conceitos) mais similares possíveis em duas línguas diferentes. Esse método realiza expansão-e-contração para encontrar o conceito equivalente (possível tradução) da seguinte maneira:

- primeiramente, é expandido o conceito semente (fornecido pelo léxico) para sua vizinhança de nós e links gerando uma sub-rede originada do nó/link dado com peso diferente;
- o contexto do nó/link dado é encontrado na linguagem alvo baseado nas conexões existentes, isso infere na sub-rede equivalente na linguagem alvo (suposta tradução de contexto) e pontua a correlação de cada nó/link da sub-rede alvo;

[4] NATools é um conjunto de ferramentas desenvolvidas para trabalhar com corpora paralelos, que está disponível livremente em: http://linguateca.di.uminho.pt/natools/ sob as especificação da GNU (General Public License).
[5] http://fjoch.com/GIZA++.html

- a contração da sub-rede para o nó/link alvo é realizada com base nas pontuações. Deste modo, o nó/link dado e o nó/link inferido possuem o contexto similar assim como seus usos, propriedades, localizações, mesmo que seus significados no dicionário possam ser diferentes.

Para a pontuação dos nós/links, são considerados três fatores:

1. Distância a partir do nó raiz: um nó filho é mais relacionado ao nó raiz que um nó neto, já que o nó neto se relaciona com o nó raiz através de uma relação de um nó filho. Portanto, quanto mais próximo um nó do nó raiz, maior sua pontuação;

2. Número de filhos dos nós: no sistema da ConceptNet, quanto maior o número de filhos dos nós de uma conexão, mais fraca é esta conexão. Por exemplo, o nó “heat” e um de seus 12 nós filhos “CapableOf – cause fire” tem uma conexão mais forte que o nó “person” e um de seus 3000 filhos “CapableOf – build”. Assim, quanto menor o número de filhos dos nós de uma conexão, mais forte será essa conexão – mais pontos para os nós relacionados;

3. Tipos das relações: todos os nós das ConceptNets estão conectados com outros nós através das relações de Minsky (1986), sendo algumas com ligações mais fortes que outras. Por exemplo, os nós “apple” e “fruit” que são conectados por “IsA” tem uma conexão mais forte que os nós “dog” e “run” que são conectados por “DesireOf”. Deste modo, conexões mais fortes terão mais pontos.

Assim, com esse processo, os conceitos mais similares terão maior pontuação e serão alinhados. Os modos de alinhamento propostos com base no estudo da estrutura das ConceptNets e dos artigos citados anteriormente, são descritos na próxima seção.

3 Recursos Linguísticos

Os experimentos descritos neste relatório foram realizados utilizando dois tipos de recursos linguísticos: um léxico bilíngue estatístico e as ConpentNets paralelas.

O léxico foi gerado automaticamente pela ferramenta NATools a partir de um *corpus* paralelo composto de artigos da versão online da revista científica brasileira Pesquisa FAPESP[6] escrita em português brasileiro (versão original) e inglês (versão traduzida). Esse corpus possui 17.397 pares de sentenças e mais de 1 milhão de *tokens* (494.391 em português e 532.121 em inglês) (CASELI, 2007).

O léxico bilíngue gerado a partir desse *corpus* contém 31.333 entradas no sentido português-inglês e 24.185 entradas no sentido inglês-português. Para cada palavra em uma entrada, o léxico apresenta até 10 possíveis traduções pontuadas estatisticamente com base em suas frequências de coocorrência no *corpus*. Assim, cada uma das 10 possíveis traduções é acompanhada de sua probabilidade de tradução. A Figura 1 a seguir apresenta um extrato desse léxico com uma entrada para a palavra em português "realidade" seguida de uma entrada em inglês para a palavra "*reality*". De acordo com as informações apresentadas nessas entradas, é possível constatar que a palavra em português ocorreu 49 vezes no *corpus* (valor que segue `count`) enquanto a palavra em inglês, 48. Além disso, verifica-se que o par (realidade, *reality*) é a melhor tradução com mais de 91% de probabilidade (calculada automaticamente por NATools com base na coocorrência desse par no *corpus* paralelo).

```
"realidade" => {
    count => 49,
    trans => {
        "actually" => 0.0812556222081184,
        "reality" => 0.918744444847107,
    },
},

"reality" => {
    count => 48,
    trans => {
        "realidades" => 0.0104384133592248,
        "concretizar" => 0.021004643291235,
        "realidade" => 0.957692444324493,
        "sonhos" => 0.010864470154047,
    },
},
```

Figura 1. Extrato do léxico português-inglês gerado automaticamente por NATools a partir do *corpus* paralelo da revista Pesquisa FAPESP.

[6] http://www.revistapesquisa.fapesp.br/

Vale dizer, ainda, que o conjunto de possíveis pares de traduções representados nesse léxico não é o mesmo produzido pelo alinhamento de palavras (lexical) gerado automaticamente por ferramentas como o GIZA++ já que, no léxico, várias possibilidades de tradução são apresentadas e não apenas a melhor no contexto das sentenças paralelas (como ocorre no alinhamento lexical). As correspondências bilíngues e suas probabilidades presentes nesse léxico estão armazenadas em um arquivo texto UTF-8 no formato apresentado na Figura 1. Tal arquivo, denominado `lexico.txt` nos comandos apresentados neste documento, é passado como parâmetro de entrada para o script que implementa os modos de alinhamento uma vez que seu conteúdo constitui a base quase todos.

Além do léxico bilíngue português-inglês gerado por NATools, outro recurso (passado como parâmetro de entrada) obrigatório utilizado nesse trabalho são as duas redes semânticas de conceitos (ConceptNets). As ConceptNets armazenam conhecimento de senso comum por meio de uma coleção de nós (representando os conceitos) que são conectados por arcos (representando as relações entre dois conceitos). As relações são definidas com base em estudos da teoria de Minsky (1986) de como funciona o pensamento humano.

Como já mencionado anteriormente, as ConceptNets utilizadas neste trabalho foram derivadas dos projetos *Open Mind Common Sense* nos EUA (OMCS) (SINGH, 2002) e no Brasil (OMCS-BR) (ANACLETO et al., 2008). A ConceptNet brasileira (CN-BR) compartilha 17 das relações com a norte-americana (CN-EN), como CapableOf, UsedFor, IsA, PartOf, DefinedAs, MadeOf entre outras, com algumas ligeiras variações de nomenclatura devidamente tratadas nesse trabalho como descrito a seguir. O número de relações e conceitos presentes em cada uma dessas ConceptNets é detalhado na Tabela 1 e extratos da CN-BR e da CN-EN são apresentados nas Figuras 2 e 3.

Tabela 1. Quantidade de relações e conceitos presentes em cada uma das ConceptNets utilizadas nesse projeto: ConceptNet brasileira (CN-BR) e ConceptNet americana (CN-EN).

Quantidade de	**CN-BR**	**CN-EN**
Relações (arcos)	138.868	533.136
Conceitos (nós)	63.929	276.859

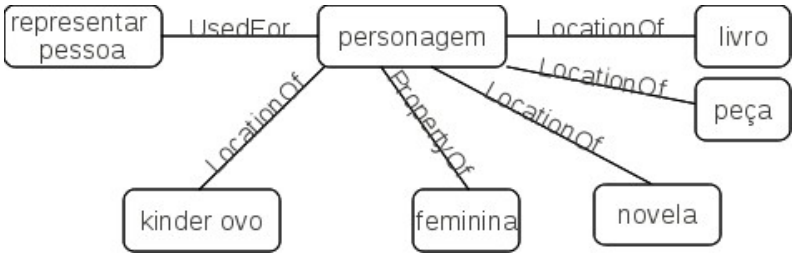

Figura 2. Extrato da CN-BR.

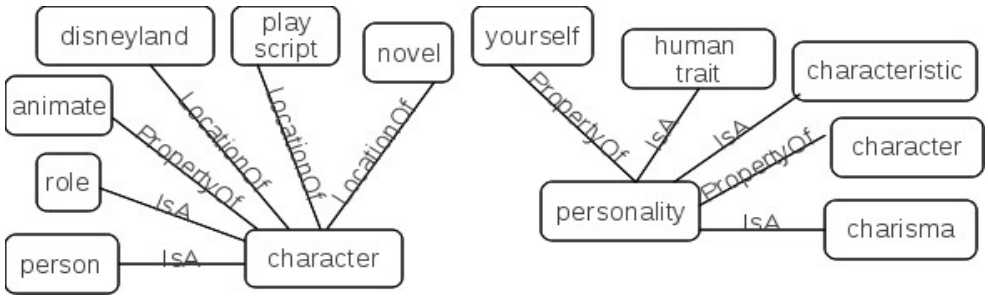

Figura 3. Extrato da CN-EN.

O conteúdo das ConceptNets, representado graficamente nas Figuras 2 e 3, está, na verdade, armazenado em arquivos texto UTF-8 como apresentado nos extrato das Figuras 4 e 5, respectivamente. Tais arquivos contendo as redes brasileira e americana também são passados como parâmetro de entrada para o script que implementa os modos de alinhamento e serão referenciados neste documento sob as nomenclaturas de `CN-BR.txt` e `CN-EN.txt`, respectivamente.

```
(LocationOf "personagem" "livro infantil" "f=1;i=0")
(LocationOf "personagem" "peça" "f=1;i=0")
(LocationOf "personagem" "livro" "f=1;i=0")
(IsA "coelho da páscoa" "personagem" "f=1;i=0")
(DefinedAs "pato donald" "personagem" "f=1;i=0")
(PropertyOf "personagem" "dentuça" "f=1;i=0")
(CapableOfReceivingAction "personagem" "ser" "f=0;i=3")
(UsedFor "personagem" "representar pessoa" "f=1;i=0")
(LocationOf "personagem" "novela" "f=1;i=0")
(CapableOfReceivingAction "personagem" "caracterizar" "f=0;i=1")
(PropertyOf "personagem" "feminina" "f=0;i=1")
(PropertyOf "personagem" "laranja" "f=1;i=0")
(PartOf "romance" "personagem" "f=1;i=0")
(LocationOf "personagem" "kinder ovo" "f=1;i=0")
…
```

Figura 4. Extrato do arquivo texto (`CN-BR.txt`) que contém os conceitos e as relações apresentados na Figura 2 nas CN brasileira.

```
(IsA "character" "person" "f=5")
(IsA "character" "role" "f=5")
(HasProperty "character" "animate" "f=5")
(AtLocation "character" "disneyland" "f=5")
(AtLocation "character" "play script" "f=5")
(AtLocation "character" "novel" "f=5")
(HasProperty "character" "novel" "f=5")
…
(HasA "personality" "many attribute" "f=5")
(HasProperty "personality" "who be" "f=5")
(IsA "personality" "charisma" "f=5")
(IsA "personality" "be" "f=5")
(IsA "personality" "trait" "f=5")
(HasProperty "personality" "spunky dull" "f=5")
(HasA "personality" "standardize test" "f=5")
(IsA "personality" "character trait" "f=5")
(HasA "personality" "swagger" "f=5")
(HasProperty "personality" "who person" "f=5")
(HasProperty "alias" "personality" "f=5")
(HasProperty "personality" "yourself" "f=5")
(HasProperty "personality" "characteristic" "f=5")
(IsA "personality" "characteristic" "f=5")
(HasA "personality" "many type" "f=5")
(HasProperty "personality" "demeanour" "f=5")
(IsA "personality" "human trait" "f=5")
(IsA "personality" "internal feeling" "f=5")
(HasProperty "personality" "character" "f=5")
(HasA "personality" "character" "f=5")
(HasA "inanimate object" "personality" "f=-5")
(HasPrerequisite "flirt" "personality" "f=5")
(HasPrerequisite "groom" "personality" "f=5")
(IsA "flair" "personality" "f=5")
(HasProperty "character" "personality" "f=5")
(IsA "brat" "personality" "f=5")
(IsA "character" "personality" "f=5")
(HasA "character" "personality" "f=5")
(IsA "disposition" "personality" "f=5")
…
```

Figura 5. Extrato do arquivo texto (`CN-EN.txt`) que contém os conceitos e as relações apresentados na Figura 3 nas CN americana.

A partir dos extratos apresentados nas Figuras 4 e 5 é possível notar que há relações com nomes diferentes, mas com mesmo significado. Por exemplo, a relação AtLocation na CN-EN é equivalente à relação LocationOf na CN-BR. Para que esses casos fossem tratados da maneira adequada, ou seja, que pares de relações equivalentes fossem tratados como se tivessem o mesmo nome, optou-se por adotar a nomenclatura da CN-BR substituindo os nomes das relações na CN-EN pelos correspondentes na CN-BR conforme apresentado na Tabela 2.

Tabela 2. Mapeamento da nomenclatura das relações para garantir a equivalência. Os nomes das relações na CN-EN apresentadas na primeira coluna foram substituídos pelos nomes equivalentes na CN-BR apresentados na segunda coluna.

CN-EN	**CN-BR**
AtLocation	LocationOf
Causes	EffectOf
CausesDesire	DesirousEffectOf
ReceivesAction	CapableOfReceivingAction
Desires	DesireOf
HasFirstSubevent	FirstSubeventOf
HasLastSubevent	LastSubeventOf
HasPrerequisite	PrerequisiteEventOf
HasProperty	PropertyOf
HasSubevent	SubeventOf
MotivatedByGoal	MotivationOf

Por fim, verificou-se quantas entradas existem em comum entre o léxico bilíngue e as ConceptNets fonte e alvo, ou seja, a intersecção entre os conceitos presentes nas redes semânticas e as palavras para as quais existem entradas no léxico bilíngue. Esses valores são apresentados na Tabela 3. O resultado dessa verificação demonstra que a intersecção entre os recursos linguísticos é pequena. Como o léxico bilíngue é o recurso principal da estratégia adotada nesse trabalho para o alinhamento das ConceptNets, essa pequena intersecção acarreta uma baixa cobertura dos modos propostos, ou seja, muitos nós permanecerão não-alinhados nas ConceptNets fonte e alvo porque os conceitos que representam não ocorrem no léxico bilíngue usado como base para o alinhamento. Mais precisamente, a cobertura dos modos de alinhamento baseados no léxico está limitada a apenas 8,07% (5.160 de 63.929) dos conceitos representados na CN-BR e apenas 2,67% (7.392 de 276.859) dos conceitos representados na CN-EN.

Tabela 3. Quantidade de palavras no léxico e conceitos nas ConceptNets utilizadas nesse projeto, bem como a intersecção entre eles.

Recurso	**Português**	**Inglês**
Léxico bilíngue	31.333	24.185
ConceptNet	63.929	276.859
Intersecção	5.160	7.392

A partir dos recursos aqui descritos, a próxima seção apresenta os algoritmos propostos e implementados no intuito de se alcançar o objetivo desse trabalho: o alinhamento de ConceptNets paralelas, ou seja, o mapeamento entre os nós de uma ConceptNet fonte e os nós de outra ConceptNet (alvo) em idioma distinto, mas com conceitos que são a tradução dos conceitos fonte.

4 Alinhamento de ConceptNets

A proposta adotada nesse trabalho para se determinar o alinhamento entre ConceptNets paralelas foi explorar basicamente dois critérios de alinhamento: (1) as correspondências entre as palavras que representam os conceitos em cada nó, obtidas por meio de consultas ao léxico bilíngue e (2) a estrutura hierárquica dos conceitos em termos da quantidade de filhos que possuem e/ou as relações em que estão envolvidos.

Assim, para realizar o alinhamento de nós em duas ConceptNets diferentes, foram definidos cinco modos de alinhamento nomeados como: A, B1, B2, B3 e C. Porém, antes de explicar cada modo de alinhamento separadamente, alguns esclarecimentos devem ser apresentados. Primeiramente, cada nó fonte (para o qual o alinhamento está sendo feito) é chamado de nó raiz e a sub-rede que contém apenas esse nó raiz é uma rede de nível 0. Todos os outros nós que se conectam com o nó raiz são chamados de nós filhos e as relações (*links*) entre os nós filhos e o nó raiz são denominadas relações *links* filhos. Por fim, a sub-rede composta pelo nó raiz, nós filhos e *links* filhos é uma rede de nível 1.

4.1 Modo de alinhamento A

O alinhamento é realizado com base somente no léxico bilíngue. Nesse processo, cada nó (nó raiz) da ConceptNet fonte é buscado no léxico bilíngue e, se encontrado, buscam-se as traduções para esse nó raiz na ConceptNet alvo, em ordem decrescente de probabilidade (da maior probabilidade de tradução para a menor), alinhando o nó raiz com a primeira melhor tradução encontrada. Vale mencionar que, na versão atual do modo A, um limite mínimo de probabilidade (`<limite>`) pode ser usado para filtrar traduções pouco frequentes que provavelmente representam casos espúrios de tradução.

Por fim, se nenhuma das traduções oferecidas pelo léxico estiver presente na ConceptNet alvo, nenhum alinhamento é estabelecido para o nó raiz em questão. A Figura 6 a seguir apresenta o algoritmo para esse modo de alinhamento o qual pode ser executado, por exemplo, por meio do comando:

```
perl alinhamento_ConceptNets_v9.pl
      CN-BR.txt
      CN-EN.txt
      lexico.txt
      A
      -inter
```

que tem como parâmetros de entrada:

- `CN-BR.txt` - o arquivo contendo a ConceptNet brasileira;
- `CN-EN.txt` - o arquivo contendo a ConceptNet americana;
- `lexico.txt` - o arquivo contendo o léxico bilíngue;
- `A` - a indicação do modo de alinhamento a ser utilizado;
- `-inter` - a indicação de que o alinhamento será feito nos sentidos fonte-alvo e alvo-fonte e o resultado será o conjunto de alinhamentos propostos em ambos os sentidos (a intersecção).

```
modo de alinhamento A
início
   para cada nó da ConceptNet fonte faça
   |  se o nó está definido no léxico bilíngue então
   |  |  ordena traduções em ordem decrescente de probabilidade
   |  |  para cada tradução (ordem decrescente de probabilidade) deste nó faça
   |  |  |  se a probabilidade da tradução for maior que <limite> então
   |  |  |  |  se tradução está definida na ConceptNet alvo então
   |  |  |  |     alinha nó fonte com esta tradução
   |  |  |  |  fim-se
   |  |  |  fim-se
   |  |  fim-para
   |  fim-se
   fim-para
fim A
```

Figura 6. Algoritmo para o modo de alinhamento A.

O modo A tem, ainda, a possibilidade de ser executado com a especificação de um valor mínimo de probabilidade de tradução a ser considerado, ou seja, apenas traduções que igualem ou superem tal valor serão consideradas como possíveis candidatas ao alinhamento. Nesse caso, o comando para execução do modo A com, por exemplo, um limite mínimo de probabilidade igual a 40% seria:

```
perl alinhamento_ConceptNets_v9.pl
      CN-BR.txt
      CN-EN.txt
      lexico.txt
      A
      -inter
      0.4
```

4.2 Modos de alinhamento B1, B2 e B3

Os três modos de alinhamento nomeados com a inicial B seguem uma estratégia um pouco distinta para determinar o alinhamento: com base no método expande-e-contrai (CHUNG et al., 2007). De acordo com esse método:

1. O contexto do nó sendo alinhado (nó raiz) é procurado expandindo seu conceito para seus *links* (conexões) vizinhos e gerando uma sub-rede de nível 1 originada no nó raiz;
2. O contexto do nó dado é procurado na ConceptNet alvo considerando-se as conexões existentes, isso infere na sub-rede correspondente na linguagem alvo e pontua a correlação de cada nó e *link* de uma sub-rede alvo;
3. A sub-rede alvo é, então, contraída para o nó alvo com base na pontuação previamente calculada.

Desse modo, o nó dado e o nó inferido terão um contexto similar assim como seus usos, propriedades, ou locais, mesmo que possuam significados não relacionados no léxico.

4.2.1 Modo de alinhamento B1

No modo de alinhamento B1, como no modo A, o nó raiz é procurado no léxico bilíngue e, se encontrado, cada tradução do nó raiz é buscada na ConceptNet alvo. Entretanto, diferentemente de A, os filhos do nó dado e os filhos de cada nó tradução são contados e o alinhamento do nó dado é estabelecido com o nó tradução que possua o menor número de filhos (novamente, se nenhuma das traduções existir na ConceptNet alvo, não haverá alinhamento). A escolha por utilizar um critério de alinhamento baseado no número de filhos se baseou no trabalho de Chung et alli (2007), onde um dos fatores do sistema de pontuação de é o número de filhos que o nó possui. Além disso, de acordo com (SINGH, 2002), quanto maior o número de filhos dos nós de uma conexão na ConceptNet, mais fraca é essa conexão, isto é, a força da conexão (*link*) é afetada pelo número de filhos dos nós envolvidos. A Figura 7 exibe o algoritmo para o modo de alinhamento B1, o qual poderia ser executado, por exemplo, por meio do comando:

```
perl alinhamento_ConceptNets_v9.pl
      CN-BR.txt
      CN-EN.txt
      lexico.txt
      B1
      -inter
```

```
modo de alinhamento B1
início
   para cada nó da ConceptNet fonte faça
   |  se o nó está definido no léxico bilíngue então
   |  |  para cada tradução deste nó faça
   |  |  |  se tradução está definida na ConceptNet alvo então
   |  |  |     conta e armazena o número de filhos da tradução
   |  |  |  fim-se
   |  |  fim-para
   |  |  alinha nó fonte com o nó tradução que possua o menor número de filhos
   |  fim-se
   fim-para
fim B1
```

Figura 7. Algoritmo para o modo de alinhamento B1.

4.2.2 Modo de alinhamento B2

Similar ao modo B1, no modo B2, o nó raiz é procurado no léxico bilíngue e suas traduções, na ConceptNet alvo. No entanto, em B2, o critério de alinhamento aplicado é a correlação entre relações fonte e alvo ao invés de número de filhos, como em B1. A sub-rede com a maior pontuação calculada com base nas relações de Minsky que ela contém é contraída e o alinhamento entre o nó raiz e o nó contraído é realizado (mais uma vez, se nenhuma das traduções existir na ConceptNet alvo, não ocorre o alinhamento).

O cálculo da pontuação de uma sub-rede é realizado com base nas relações de Minsky que ela contém pontuando tais relações da seguinte maneira:

- IsA e DefinedAs são pontuadas com +4;
- MadeOf, PropertyOf e PartOf são pontuadas com +2;
- CapableOf, LocationOf, EffectOf, DesirousEffectOf, UsedFor e CapableOfReceivingAction são pontuadas com +1.

Esses valores foram definidos empiricamente, levando-se em conta a força do significado de cada relação. Por exemplo, relações que definem os conceitos como IsA e DefinedAs têm maior pontuação do que relações mais genéricas como LocationOf e UsedFor. As demais relações não citadas anteriormente não são pontuadas na versão atual do modo de alinhamento B2. A Figura 8 apresenta o algoritmo para o modo de alinhamento B2 o qual poderia ser executado, por exemplo, por meio do comando:

```
perl alinhamento_ConceptNets_v9.pl
      CN-BR.txt
      CN-EN.txt
      lexico.txt
      B2
      -inter
```

```
modo de alinhamento B2
início
   para cada nó da ConceptNet fonte faça
   |  se o nó está definido no léxico bilíngue então
   |  |  para cada tradução deste nó faça
   |  |  |  se tradução está definida na ConceptNet alvo então
   |  |  |  |  para todas as relações do nó raiz
   |  |  |  |  |  para todas as relações da tradução
   |  |  |  |  |  |  se a relação fonte é igual/equivalente à relação alvo então
   |  |  |  |  |  |     pontua tradução conforme relação
   |  |  |  |  |  |  fim-se
   |  |  |  |  |  fim-para
   |  |  |  |  fim-para
   |  |  |  fim-se
   |  |  fim-para
   |  |  alinha nó fonte com tradução de maior pontuação
   |  fim-se
   fim-para
fim B2
```

Figura 8. Algoritmo para o modo de alinhamento B2.

4.2.3 Modo de alinhamento B3

O modo de alinhamento B3 faz a junção dos modos B1 e B2 da seguinte maneira. O nó raiz é procurado no léxico bilíngue e, se encontrado, cada tradução do nó raiz é buscada na ConceptNet alvo. Para cada tradução encontrada na ConceptNet alvo, as relações do nó dado e do nó tradução são pontuadas da mesma maneira que no modo B2, porém, em B3, uma pontuação também é atribuída as traduções de acordo com o número de filhos que elas contém. Nesse caso, atribui-se pontuação máxima igual a 4 (mesmo peso atribuído às relações mais fortes) para a tradução com o menor número de filhos e assim por diante até a tradução com o maior número de filhos. Desse modo, a estratégia de privilegiar traduções com menor número de filhos do modo B1 é mantida em B3.

Por fim, a sub-rede com a maior pontuação final (calculada com base nas relações e n o número de filhos) é contraída e o alinhamento entre o nó raiz e o nó contraído é realizado

(outra vez, não haverá alinhamento se nenhuma das traduções for encontrada na ConceptNet alvo). A Figura 9 apresenta o algoritmo para o modo de alinhamento B3 o qual, equivalentemente aos modos anteriores, poderia ser executado por meio do comando:

```
perl alinhamento_ConceptNets_v9.pl
      CN-BR.txt
      CN-EN.txt
      lexico.txt
      B3
      -inter
```

```
modo de alinhamento B3
início
   para cada nó da ConceptNet fonte faça
   |  se o nó está definido no léxico bilíngue então
   |  |  para cada tradução deste nó faça
   |  |  |  se tradução está definida na ConceptNet alvo então
   |  |  |  |  conta e armazena o número de filhos da tradução       {como em B1}
   |  |  |  |  para todas as relações do nó raiz                     {como em B2}
   |  |  |  |  |  para todas as relações da tradução
   |  |  |  |  |  |  se a relação fonte é igual/equivalente à relação alvo então
   |  |  |  |  |  |     pontua tradução conforme relação
   |  |  |  |  |  |  fim-se
   |  |  |  |  |  fim-para
   |  |  |  |  fim-para
   |  |  |  fim-se
   |  |  fim-para
   |  |  pontua tradução quanto ao número de filhos
   |  |  alinha nó fonte com tradução de maior pontuação
   |  fim-se
   fim-para
fim B3
```

Figura 9. Algoritmo para o modo de alinhamento B3.

4.3 Modo de alinhamento C

Todos os modos descritos até aqui alinham os nós de duas ConceptNets apenas se a palavra representada em um determinado nó for encontrada no léxico bilíngue. Dessa maneira, como descrito na seção 3, o alinhamento apresentado até então está limitado ao máximo de 5.160 conceitos em português e 7.392 conceitos em inglês, que é a quantidade de palavras do léxico que também ocorrem nas ConceptNets (valores para a intersecção entre esses recursos apresentados na Tabela 3).

Como uma alternativa para os modos baseados no léxico bilíngue descritos até então, o modo C foi proposto com uma estratégia diferente, e bastante ingênua, de alinhar os nós restantes, ou seja, alinhar os nós que não foram alinhados por qualquer um dos modos anteriormente descritos (A, B1, B2 e B3) ou, possivelmente, a combinação de todos eles. Para tanto, o modo C funciona da seguinte maneira. Ele recebe como parâmetros de entrada os alinhamentos base (fonte-alvo e alvo-fonte) gerados por outro modo de alinhamento ou, possivelmente, a junção de todos os modos de alinhamento anteriormente descritos.

A partir desse alinhamento base representado como o conjunto de pares de nós alinhados (nó_fonte_alinhado, nó_alvo_alinhado), para cada par de nós alinhados, verifica-se se o nó fonte alinhado consta na ConceptNet fonte[7]. Em seguida, para cada relação desse nó, armazena-se a relação e o nó filho envolvido e verifica-se se esse nó filho está presente na ConceptNet fonte e se ele já não foi previamente alinhado pelo modo base. O mesmo é feito com o conceito alvo alinhado (nó_alvo_alinhado), ou seja verifica-se sua presença na ConceptNet alvo, armazena-se a relação e o nó filho, verifica-se se esse nó filho alvo está presente na ConceptNet alvo e se ainda se mantém não-alinhado. Por fim, verifica-se se a relação que o nó fonte alinhado mantém com o nó filho fonte em questão é a mesma que o nó alvo alinhado mantém com seu nó filho. Em caso afirmativo, o nó filho alvo é considerado um candidato ao alinhamento com o nó filho fonte. O alinhamento final só será estabelecido se houver apenas um candidato ao alinhamento que satisfaça os critérios especificados acima. A Figura 10 apresenta o algoritmo para o modo de alinhamento C que tem um comando de execução um pouco distinto dos anteriores, pois não recebe como entrada o léxico bilíngue, mas sim dois arquivos com o resultado do alinhamento produzido anteriormente por um ou mais modos de alinhamento em cada sentido (fonte-alvo e alvo fonte), como apresentado no comando:

```
perl alinhamento_ConceptNets_v9.pl
      CN-BR.txt
      CN-EN.txt
      C
      alinhamento_base_fonte-alvo.txt
      alinhamento_base_alvo-fonte.txt
      -inter
```

[7] No modo C, a verificação de pertinência de um conceito já alinhado previamente na ConceptNet para a qual se realiza o alinhamento no momento é necessária porque tal alinhamento base pode ter sido gerado a partir de outras versões das ConceptNets o que gera a possibilidade de o conceito anteriormente existente e alinhado já não mais existir na versão atual da ConceptNet.

```
modo de alinhamento C
início
   para cada nó do alinhamento base (nó_fonte_alinhado, nó_alvo_alinhado) faça
   |  se o nó_fonte_alinhado está definido na ConceptNet fonte então
   |  |  para todas as relações desse nó_fonte_alinhado faça
   |  |  |  armazena relação em relação_fonte e filho em nó_filho_fonte
   |  |  |  se ainda não há alinhamento feito para esse nó_filho_fonte então
   |  |  |  |  se o nó_alvo_alinhado está na ConceptNet alvo então
   |  |  |  |  |  para todas as relações do nó_alvo_alinhado faça
   |  |  |  |  |  |  armazena relação em relação_alvo e filho em nó_filho_alvo
   |  |  |  |  |  |  se relação_fonte = relação_alvo então
   |  |  |  |  |  |     nó_filho_alvo é candidato a alinhamento com nó_filho_fonte
   |  |  |  |  |  |  fim-se
   |  |  |  |  |  fim-para
   |  |  |  |  fim-se
   |  |  |  fim-se
   |  |  fim-para
   |  |  se houver apenas um candidato a alinhamento com nó_filho_fonte então
   |  |     alinha nó_filho_fonte com o único candidato
   |  |  fim-se
   |  fim-se
   fim-para
fim C
```

Figura 10. Algoritmo para o modo de alinhamento C.

A Figura 11 a seguir resume graficamente a relação dos modos de alinhamento descritos anteriormente com os critérios usados no alinhamento:

- dic – alinhamento feito com base no léxico bilíngue,
- # - alinhamento feito baseado no número de filhos,
- * - alinhamento feito baseado nas relações.

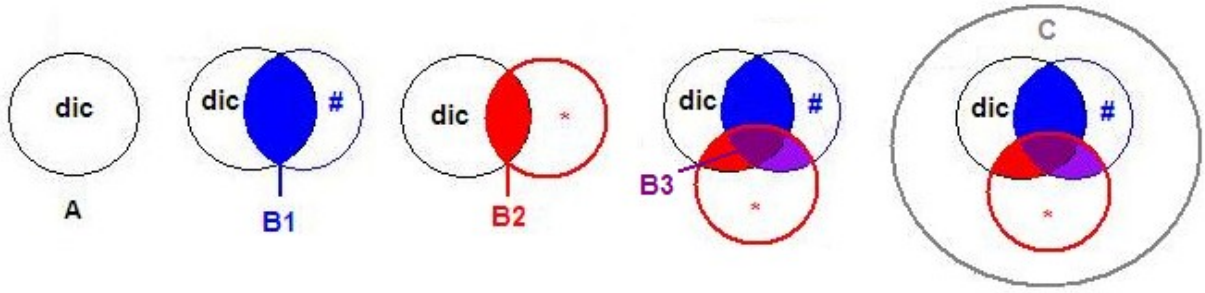

Figura 11. Representação gráfica dos modos de alinhamento e os critérios usados por cada um.

Interpretando os gráficos da Figura 11, tem-se um resumo dos modos de alinhamento:

- modo A – alinha apenas com base no léxico bilíngue (dic);
- modo B1 – alinha com base no léxico (dic) e no número de filhos (#);
- modo B2 – alinha com base no léxico (dic) e nas relações entre os conceitos (*);
- modo B3 – alinha com base no dicionário (dic), no número de filhos (#) e, na relação entre os conceitos (*). Nesse caso, veja que B3 é a intersecção de B1 e B2 uma vez que utiliza critérios de alinhamento adotados por cada um desses modos;
- modo C – alinha os nós que não foram alinhados pelos outros modos. Veja que, nesse caso, o modo C é indicado como o complemento do que os outros modos de alinhamento conseguem alinhar já que ele é o modo de alinhamento que alinha nós não alinhados pelos outros modos.

5 Experimentos e Resultados

Os modos de alinhamento descritos na seção anterior foram avaliados resultando nos valores apresentados a seguir. Para tanto, cada modo foi executado por meio dos comandos exemplificados na seção 4, tendo como entrada os recursos apresentados na seção 3 e, como saída, os alinhamentos em ambos os sentidos de tradução: fonte-alvo e alvo-fonte. Além dos alinhamentos nos dois sentidos de tradução separadamente, considerou-se também o resultado da intersecção de ambos os sentidos em todos os modos de alinhamento. Especificamente para a execução do modo C, foram gerados dois arquivos contendo a união dos alinhamentos produzidos pelos demais modos (A, B1, B2 e B3) nos dois sentidos: fonte-alvo e alvo-fonte. Tais arquivos foram usados como os alinhamentos-base para execução do modo de alinhamento C (como parâmetros de entrada).

A Tabela 4 a seguir resume a quantidade de alinhamentos gerados por cada um dos modos de alinhamento descritos anteriormente. Para o modo de alinhamento A experimentou-se duas possibilidades de alinhamento: sem limite mínimo para a probabilidade de tradução a ser considerado para as entradas do léxico bilíngue (A_0) e com o limite mínimo de 40% (A_40), ou seja, apenas as traduções com probabilidade de tradução maior ou igual a 40% nas entradas do léxico foram consideradas candidatas ao alinhamento.

Tabela 4. Quantidade de alinhamentos gerados por cada modo de alinhamento avaliado.

	A_0	**A_40**	**B1**	**B2**	**B3**	**A_40+B***	**C**
fonte-alvo	3.139	2.240	3.139	1.500	3.139	4.085	159
alvo-fonte	2.842	1.920	2.842	865	2.842	3.629	200
intersecção	2.034	1.464	1.851	406	1.823	--	10

Por essa tabela, tem-se que os modos que têm maior número de alinhamentos são A_0, B1 e B3, ou seja, o modo que não restringe a probabilidade de tradução (A_0) e os modos que realizam o alinhamento com base no número de filhos como o único (B1) ou um de seus critérios de alinhamento (B3). O modo A_40 gerou um número menor de alinhamentos do que os modos citados por causa de sua restrição de probabilidade de tradução (do léxico) ser de, ao menos, 40%. Os alinhamentos do modo C foram gerados como aqueles que não foram gerados pelos modos A_40, B1, B2 e B3, ou seja, a união dos alinhamentos gerados por esses modos, nos sentidos fonte-alvo e alvo-fonte, foi passada como parâmetro de entrada para o modo C. Assim, o modo C alinhou os conceitos excluindo-se os 4.085 alinhamentos fonte-

alvo e os 3.629 alinhamentos alvo-fonte gerados por A_40 e os modos B*, resultando na menor quantidade de alinhamentos de todos os modos investigados, como era de se esperar.

Para proceder com a avaliação dos modos, considerou-se apenas o resultado da intersecção dos alinhamentos gerados, por cada modo, nos dois sentidos (fonte-alvo e alvo-fonte), ou seja, as quantidades de alinhamentos apresentadas na última linha da Tabela 4. Acredita-se que ao considerar o arquivo de saída com a intersecção dos alinhamentos há uma maior probabilidade de esses alinhamentos estarem corretos, pois foram encontrados nos dois sentidos da tradução (fonte-alvo e alvo-fonte).

Considerando-se todos esses arquivos de intersecção gerados pelos cinco modos, 2.840 alinhamentos diferentes foram produzidos envolvendo um total de 2.524 conceitos em português e 2.445 conceitos em inglês. Assim, a cobertura obtida em relação à quantidade de conceitos representados nas CNs brasileira e americana foi, respectivamente, de apenas 3,95% (2.524 de 63.929 conceitos na CN-BR) e 0,88% (2.445 de 276.859 conceitos na CN-EN). A baixa cobertura já era esperada uma vez que, como descrito na seção 3, o número máximo de nós que poderiam ser alinhados está limitado à intersecção entre as palavras presentes nos dois recursos utilizados no alinhamento (léxico bilíngue e ConceptNets) que é de 5.160 palavras em português (8,07% = 5.160 de 63.929) e 7.392 palavras em inglês (2,67% = 7.392 de 276.859). Portanto, é possível concluir que alcançou-se para CN-BR e CN-EN, respectivamente, 48,91% (2.524 de 5.160 palavras na intersecção de léxico bilíngue com CN-BR) e 33,08% (2.445 de 7.392 palavras na intersecção de léxico bilíngue com CN-BR) da cobertura máxima que se poderia alcançar com a estratégia de alinhamento proposta com base no léxico bilíngue.

A avaliação dos alinhamentos gerados foi, então, realizada para uma amostra contendo 30% do total de 2.840 alinhamentos diferentes gerados, ou seja, 852 alinhamentos foram avaliados manualmente. A quantidade (e respectiva porcentagem) de entradas de cada modo avaliada manualmente é apresentada na Tabela 5.

Tabela 5. Quantidade de alinhamentos avaliados manualmente em relação à quantidade gerada por cada modo de alinhamento (Tabela 4).

	A_0	**A_40**	**B1**	**B2**	**B3**	**C**	**TOTAL**
intersecção	2.034	1.464	1.851	406	1.823	10	2.840
# amostra	711	625	665	145	676	10	852
% amostra	34,95%	42,69%	35,93%	35,71%	37,08%	100,00%	30,00%

Essa amostra total de 852 alinhamentos foi dividida e avaliada manualmente por dois nativos do português com conhecimento em inglês. Nessa tarefa, cada avaliador classificou os alinhamentos como V (verdadeiro) ou F (falso) de acordo com os seguintes critérios: (i) a palavra fonte é uma possível tradução da palavra alvo (e vice-versa) e (ii) as relações fonte e alvo deixam claro que o conceito representado pela palavra fonte é o mesmo representado pela palavra alvo. Dessa forma, uma entrada foi avaliada como verdadeira se ambos critérios foram satisfeitos, como falsa, caso contrário.

Além disso, uma sobreposição de 176 entradas foi inserida na amostra para permitir a verificação da concordância entre avaliadores por meio do cálculo da medida kappa (CARLETTA, 1996) usando o script Perl `kappaDiagnosis.pl`. Para tanto, primeiro foi necessário converter os arquivos com as avaliações dos juízes humanos para o formato de entrada desse script, tal conversão foi realizada pelo script `converte_alinhamento_para_kappa.pl` como mostram os comandos a seguir:

```
perl converte_alinhamento_para_kappa.pl
      avaliacao_EM_COMUM.txt
      avaliacao_Amostra1.txt
      avaliacao_Amostra2.txt > entrada_kappa.txt
```

o qual recebe com entrada três arquivos:

- `avaliacao_EM_COMUM.txt` - arquivo que contém as 176 entradas avaliadas por ambos os juízes;
- `avaliacao_Amostra1.txt` e `avaliacao_Amostra2.txt` – arquivos com as classificações atribuídas por cada juiz para todas as entradas avaliadas.

e gera como saída um arquivo (`entrada_kappa.txt`) contendo as avaliações de ambos os juízes para as 176 entradas, no formato requisitado pelo script `kappaDiagnosis.pl`.

Após a conversão dos arquivos com as avaliações dos juízes humanos para o formato de entrada de `kappaDiagnosis.pl`, tal script foi executado para verificar o valor de kappa nas avaliações dos alinhamentos por meio do comando:

```
perl kappaDiagnosis.pl entrada_kappa.txt > saida_kappa.txt
```

Como resultado, os valores de kappa para as classificações dos alinhamentos foi 0,75 o que indica uma boa concordância entre os avaliadores. De acordo com Carletta (1996), um valor de k > 0,8 indica uma boa replicabilidade enquanto valores entre 0,67 e 0,8 permitem que conclusões sejam tiradas. Com base nesses valores concluímos que a diferença nas avaliações dos juízes humanos não é tão grande a ponto de impedir que conclusões sejam tiradas desse experimento.

Assim, a quantidade de alinhamentos classificados como V ou F pelos avaliadores, para as entradas produzidas por cada um dos modos de alinhamento investigados é apresentada na Tabela 6. Veja que devido ao pequeno número de entradas geradas na intersecção dos alinhamentos fonte e alvo gerados pelo modo C (apenas 10), todas foram avaliadas manualmente e classificadas como F demostrando que tal modo "ingênuo" de alinhamento não é o mais indicado para experimentos futuros.

Tabela 6. Quantidade (#) e porcentagem (%) de alinhamentos classificados como V ou F por pelo menos um dos dois juízes humanos.

	A_0	**A_40**	**B1**	**B2**	**B3**	**C**	**TOTAL**
# V	405	379	313	97	331	0	425
%	56,96%	60,64%	47,07%	66,90%	48,96%	0,00%	49,88%
# F	321	262	362	55	356	10	441
%	45,15%	41,92%	54,44%	37,93%	52,66%	100,00%	51,76%
discordância	15	16	10	7	11	0	14
%	2,11%	2,56%	1,50%	4,83%	1,63%	0,00%	1,64%
TOTAL	711	625	665	145	676	10	852

A partir dos valores da Tabela 6 é possível notar, novamente, a boa concordância entre anotadores, sendo que a discordância máxima foi verificada na avaliação dos alinhamentos gerados pelo modo B2 (4,83% das entradas foram avaliadas de modo diferente pelos dois avaliadores). Essa discrepância de avaliação maior nas entradas do modo B2 não impede a conclusão de que ele demonstrou o melhor desempenho, com cerca de 67% de seus alinhamentos classificados como verdadeiros por pelo menos um dos juízes humanos. O segundo modo de maior desempenho (60% de entradas classificadas como V) foi o A quando um limite mínimo de 40% (A_40) para a probabilidade de tradução foi considerado. Como era esperado, o método de pior desempenho foi o modo C (100% das 10 entradas avaliadas classificadas como falsas) mostrando que sua estratégia "ingênua" de alinhamento é bastante limitada.

Embora apenas cerca de 50% do total de entradas avaliadas tenham sido classificadas como verdadeiras, verificou-se que esse desempenho está fortemente relacionado a fatores externos à qualidade dos algoritmos de alinhamento. A seguir são apresentadas algumas entradas classificadas como falsas por ambos os avaliadores devido a fatores como:

- erro de etiquetação das palavras

  O caso exemplificado a seguir ilustra bem o fato de que nas CNs estão representados vários conceitos relacionados à mesma palavra. A palavra em inglês "*alter*" (que pode tanto ser verbo como substantivo) é uma possível tradução do verbo em português "alterar" indicando, nesse caso, um alinhamento possível. Porém, enquanto a CN-BR apresenta o conceito do verbo "alterar", a CN-EN representa o conceito do substantivo "*alter*" inviabilizando, assim, o alinhamento.

```
alterar <=> alter (A,B3) [F/Avaliador1] [F/Avaliador2]
alterar
(DefinedAs "alterar" "modificar" "YES")
alter
(LocationOf "alter" "church" "YES")
```

- erro de lematização das palavras

  No processo de geração das ConceptNets ocorre a lematização das palavras antes de inseri-las como nós. Nesse processo, cada palavra é reduzida a sua forma base: substantivos são substituídos pro sua versão masculina singular, verbos pela forma no infinitivo e assim por diante. Esse processo, que é realizado automaticamente, está sujeito a falhas como no caso exemplificado a seguir no qual o substantivo em português "maminha" foi substituído pelo lema "mama" resultando no alinhamento incorreto de conceitos que são, de fato, tradução mútua: mama e *breast*.

```
mama <=> breast (A,B3) [F/Avaliador1] [F/Avaliador2]
mama
(LocationOf "mama" "churrascaria" "YES")
breast
(CapableOf "breast" "fee baby" "YES")
(ConceptuallyRelatedTo "breast" "round jug" "YES")
(HasA "breast" "form regardless size" "YES")
(HasA "breast" "nipple" "YES")
(IsA "breast" "body part" "YES")
(IsA "breast" "boob" "YES")
(LocationOf "breast" "woman" "YES")
(PartOf "breast" "human" "YES")
(PartOf "breast" "torso" "YES")
(PropertyOf "breast" "enhance silicone" "YES")
(SimilarSize "breast" "melon" "YES")
(UsedFor "breast" "squeeze" "YES")
```

- erro de lematização das relações

  Da mesma maneira que um erro de lematização pode ocorrer no nó envolvido no alinhamento como apresentado anteriormente, também pode acontecer no nó envolvido na relação como é o caso da palavra “público” erroneamente lematizada como “publicar” no exemplo abaixo.

```
platéia <=> audience (B2) [F/Avaliador1] [V/Avaliador2]
platéia
(DefinedAs "platéia" "publicar" "YES")
(IsA "platéia" "publicar" "YES")
(LocationOf "platéia" "auditório" "YES")
(LocationOf "platéia" "casa de show" "YES")
(LocationOf "platéia" "circo" "YES")
(LocationOf "platéia" "concerto" "YES")
(LocationOf "platéia" "espetáculo" "YES")
(LocationOf "platéia" "ginásio de esporte" "YES")
(LocationOf "platéia" "peça" "YES")
(LocationOf "platéia" "programa de auditório" "YES")
(LocationOf "platéia" "programa de televisão" "YES")
(LocationOf "platéia" "show de rock" "YES")
(PropertyOf "platéia" "preta" "YES")
(PropertyOf "platéia" "vermelha" "YES")
(UsedFor "platéia" "exibir espetaculo" "YES")
(UsedFor "platéia" "representação teatral" "YES")
audience
(CapableOf "audience" "applaud" "YES")
(CapableOf "audience" "boo" "YES")
(CapableOf "audience" "buy ticket" "YES")
(CapableOf "audience" "crowd around singer" "YES")
(CapableOf "audience" "different size" "YES")
(CapableOf "audience" "dig band" "YES")
(CapableOf "audience" "eat" "YES")
(CapableOf "audience" "enjoy play" "YES")
(CapableOf "audience" "enjoy" "YES")
...
```

- número limitado de relações nas ConceptNets o que impede a dedução de que as palavras fonte e alvo tratam do mesmo conceito

  Outro problema que impediu que uma maior precisão fosse verificada na avaliação esteve presente em alinhamentos entre conceitos com poucas relações nas ConceptNets correspondentes. Nesses casos, as relações existentes não foram suficientes para garantir que as palavras fonte e alvo alinhadas tratavam do mesmo conceito.

```
algoritmo <=> algorithm (B3) [] [F/Avaliador2]
algoritmo
(MadeOf "algoritmo" "computador" "YES")
(MadeOf "algoritmo" "linguagem" "YES")
(MadeOf "algoritmo" "lógica" "YES")
(PartOf "algoritmo" "instrução" "YES")
algorithm
```

```
(IsA "algorithm" "process accomplish task" "YES")
(LocationOf "algorithm" "software" "YES")
(UsedFor "algorithm" "software" "YES")
(UsedFor "algorithm" "solve problem" "YES")
```

- palavras polissêmicas em apenas uma língua

  Nesse caso, os conceitos representados nos lados fonte e alvo não paralelos embora sejam possíveis traduções um do outro e isso ocorre porque as relações apresentadas em um dos lados são incompatíveis com as relações presentes no outro lado devido à polissemia do conceito.

```
assembléia <=> assembly (B3) [F/Avaliador1] [F/Avaliador2]
assembléia
(LocationOf "assembléia" "senado" "YES")
assembly
(IsA "assembly" "program language" "YES")
```

- palavras que podem ser a tradução uma da outra, porém, não no sentido apresentado nas relações

  Nesse caso, novamente o problema está nas relações que definem o conceito já que se apenas as palavras fossem consideradas, o alinhamento seria possível.

```
negro <=> black (B3) [] [F/Avaliador2]
negro
(LocationOf "negro" "senegal" "YES")
(PropertyOf "negro" "jamaica" "YES")
black
(CapableOf "black" "become white" "YES")
(CapableOfReceivingAction "black" "associate formal occasion"
"YES")
(ConceptuallyRelatedTo "black" "dark" "YES")
(ConceptuallyRelatedTo "black" "night" "YES")
(ConceptuallyRelatedTo "black" "nighttime" "YES")
(ConceptuallyRelatedTo "black" "nothing" "YES")
(ConceptuallyRelatedTo "black" "space" "YES")
(DefinedAs "black" "abscence color" "YES")
(DefinedAs "black" "absence color" "YES")
(DefinedAs "black" "absence colour" "YES")
...
```

A partir da análise dos resultados obtidos com a avaliação dos modos de alinhamento implementados nesse trabalho, algumas propostas de trabalhos futuros surgiram e serão investigadas em breve como:

- Verificar quais características das palavras que representam os conceitos levam dificultam o alinhamento. Por exemplo, será que um conceito concreto é mais fácil de ser alinhado do que um conceito abstrato?

- Implementar o modo B1 reverso, ou seja, ao invés de alinhar o candidato que possua o menor número de filhos, alinhar o candidato que possua o maior número de filhos já que o número limitado de relações foi um dos problemas levantados na avaliação manual dos alinhamentos;
- Também com o intuito de tentar melhorar o desempenho dos modos propostos, realizar o alinhamento apenas com nós que possuam um número mínimo de relações;
- Refinar o modo B2 considerando-se, além da equivalência entre relações, também a equivalência dos conceitos envolvidos nessas relações (nesse caso, por meio das traduções);
- Refinar o modo B2 não apenas relações de mesmo nome, mas também relações compatíveis como, por exemplo, IsA, DefinedAs e SimilarSize;
- Aumentar a cobertura do léxico bilíngue, ou seja, gerar um léxico maior a partir de um *corpus* paralelo maior e, assim, tentar aumentar a intersecção existente entre léxico e ConceptNets aumentando, em consequência, a cobertura dos modos de alinhamento propostos com base no léxico.

6 Considerações Finais

Esse documento apresentou um relato dos primeiros experimentos realizados no intuito de alinhar redes semânticas representando dados de senso comum em dois idiomas distintos (português do Brasil e inglês). Para tanto, considerou-se como critérios de alinhamento: as traduções presentes em um léxico bilíngue probabilístico gerado automaticamente a partir de *corpus* paralelo e a hierarquia/estrutura das redes semânticas definida pelo número de filhos e relações entre eles.

A partir do conteúdo apresentado neste relatório, conclui-se que o objetivo inicial de investigar, propor, implementar e avaliar métodos de alinhamento de conceitos nas redes semânticas sob estudo foi alcançado. Da análise derivada dos resultados verificou-se que os melhores critérios de alinhamento são a tradução e a equivalência de relações. Nesse sentido, trabalhos futuros serão desenvolvidos com o intuito de enriquecer e aprofundar os critérios de alinhamento que se mostraram mais promissores por meio da geração de um léxico bilíngue maior (a partir de novos *corpora* paralelos português-inglês) e do aprofundamento do tratamento da equivalência de relações. Em relação a esse último aprofundamento, pretende-se considerar relações compatíveis (e não apenas relações idênticas) e as traduções dos conceitos envolvidos nas relações (e não apenas os conceitos sendo alinhados) em versões futuras do modo de alinhamento B2.

7 Referências Bibliográficas

ANACLETO, J. C.; CARVALHO, A. F. P. DE ; PEREIRA, E. N. ; FERREIRA, A. M. ; CARLOS, A. F (2006) Machines with good sense: How can computers become capable of sensible reasonig? . In: BRAMER, M.. (Org.). Artificial Intelligence in Theory and Practice II – WCC 2006. 1 ed. Berlin: Springer-Verlag, 2006, v. 1, p. 195-204.

ANACLETO, J.C.; CARVALHO, A.F.P.; PEREIRA, E.N.; FERREIRA, A.M.; CARLOS, A.J.F. (2008). *Machines with good sense: How can computers become capable of sensible reasoning?* In: IFIP AI. 2008. p. 195–204.

CARLETTA, J. (1996). Assessing agreement on classification tasks: the kappa statistics. Computational Linguistics, v. 22, n. 2, p. 249–254, 1996.

CASELI, H. M. (2007). *Indução de léxicos bilíngües e regras para a tradução automática*. 158 p. Tese (Doutorado) – ICMC-USP, Abril 2007.

CHUNG, H.; LIEBERMAN, H.; BENDER, W. (2007). *GlobalMind – Bridging the Gap Between Different Cultures and Languages with Common-sense computing.* 2007.

DIAS DA SILVA, B. C.; MONTILHA, G.; RINO, L. H. M.; SPECIA, L.; NUNES, M. G. V.; OLIVEIRA JR, O. N.; MARTINS, R. T.; PARDO, T. A. S. (2007) . Introdução ao Processamento das Línguas Naturais e Algumas Aplicações . Série de Relatórios Técnicos do NILC (NILC-TR-07-10). 119 p. Agosto 2007.

DOERR, M.; IORIZZO, D. (2008). *The dream of a global knowledge network — A new approach*. ACM J. Comput. Cultur. Heritage 1, 1, Article 5 (June 2008), 23 pages.

HAVASI, C.; SPEER, R.; ALONSO, J. (2009). ConceptNet: A Lexical Resource for Common Sense Knowledge. In *Recent Advances in Natural Language Processing*. 2009.

HUTCHINS, J. (1998). Translation Technology and the Translator. In Machine Translation Review. Norfolk.

LUNDBORG, J.; MAREK, T.; METTLER M.; VOLK, M. (2007). *Using the Stockholm TreeAligner*. In Koenraad De Smedt, Jan Hajič and Sandra Kübler (Eds.): *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories.* NEALT Proceedings Series, Vol. 1 (2007), 73-78.

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

NAVIGLI, R. (2009). *Word sense disambiguation: A survey*. ACM Comput. Surv. 41, 2, Article 10 (February 2009), 69 pages. 2009.

NIRENBURG, S. (1987). Knowledge and choices in machine translation. In Machine translation – Theoretical and methodological issues, pp. 1-15. Cambridge University Press, Cambridge.

PETERS, W.; VOSSEN, P.; DÍEZ-ORZAS, P.; ADRIAENS, G. (1998). Cross-linguistic Alignment of Wordnets with an Inter-Lingual-Index. In *EuroWordNet: a multilingual database with lexical semantic networks*. p. 149 – 179. 1998.

SAMUELSSON, Y.; VOLK, M.; MAREK, T. (2008). *Human Judgements in Parallel Treebank Alignment*. In *Proceedings of the Workshop on Human Judgements in Computational Linguistics.* Manchester, United Kingdom, p. 51-57. 2008.

SAMUELSSON, Y.; VOLK, M. (2007). *Automatic Phrase Alignment: Using Statistical N-Gram Alignment for Syntactic Phrase Alignment*. In Koenraad De Smedt, Jan Hajič and

Sandra Kübler (Eds): *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories*. NEALT Proceedings Series, Vol. 1 (2007), 139-150.

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

TINSLEY, J; HEARNE, M; WAY, A. (2009). Parallel Treebanks in Phrase-Based Statistical Machine Translation. National Centre for Language Technology Dublin City University, Ireland. 2009.

TINSLEY, J.; ZHECHEV, V.; HEARNE, M.; WAY, A. (2007). Robust Language Pair-Independent Sub-Tree Alignment. In *Machine Translation Summit XI*. Copenhagen, Denmark. pp.467-474. 2007.

ZHECHEV, V. (2009). *Unsupervised Generation of Parallel Treebanks through Sub-Tree Alignment*. DORAS - Dublin City University institutional repository, Ireland, 2009. Disponível em: http://www.scientificcommons.org/54536595. 2009.

Agradecimentos

Agradecemos ao Programa Institucional de Bolsas de Iniciação Científica (PIBIC) do CNPq pelo apoio financeiro sem o qual não seria possível realizar os experimentos aqui apresentados. Agradecemos também ao mestrando Bruno Akio por seus comentários a respeito do processo de consulta do banco de dados das ConceptNets.