Aprimoramento do cálculo da similaridade semântico-estrutural: um estudo voltado a estruturas ontológicas em língua portuguesa

Josiane Fontoura dos Anjos[1,2] **Vera Lúcia Strube de Lima**[1]

[1] Pontifícia Universidade Católica do Rio Grande do Sul - PUCRS
Programa de Pós-Graduação em Ciência da Computação - PPGCC
Avenida Ipiranga, 6681 - Prédio 32 - Partenon
CEP 90619-900 Porto Alegre - RS - Brasil

[2]Universidade da Região da Campanha - URCAMP
Campus Universitário de Alegrete
Centro de Ciências da Economia e Informática
Praça Getúlio Vargas, 47 - Centro
CEP 97542-570 Alegrete - RS - Brasil

{josiane.brandolt,vera.strube}@pucrs.br

Abstract. *This paper provide strategies to improve the SiSe measure (Semantic similarity) [Freitas 2007], which aims to calculate the similarity between terms of different ontological structures. The strategies introduced improve the results of the semantic similarity coefficient, mainly with the analysis of false positives and false negatives.*

Resumo. *Este artigo fornece estratégias para aprimorar uma medida de similaridade semântica entre estruturas ontológicas, denominada SiSe proposta por Freitas [Freitas 2007]. As estratégias apresentadas visam melhorar os resultados dos coeficientes de similaridade baseados na análise dos falsos positivos e falsos negativos.*

1. Introdução

As ontologias desempenham um papel de importância crescente nas aplicações que envolvem a representação do conhecimento ou aplicações que exigem o emprego de termos precisamente definidos. No entanto, com o uso generalizado de ontologias, surgem muitos problemas. Os usuários ou engenheiros de ontologias frequentemente contam com uma ontologia principal que utilizam para navegar ou consultar dados. Mas precisam estender, adaptar ou comparar suas ontologias com outras existentes [Maedche and Staab 2002]. Para tratar esses problemas, estudos voltados à similaridade entre ontologias buscam analisar as semelhanças entre entidades de ontologias distintas e permitem levar a um processo de aproximação, denominado "mapeamento". A similaridade entre ontologias refere-se à comparação de ontologias inteiras ou das entidades que as compõem. Essa comparação retorna um valor numérico (por exemplo, um valor no intervalo [0,1] representando um percentual de 0 a 100), que indica maior ou menor grau de similaridade entre duas entidades ou ontologias [Ehrig 2007].

A similaridade costuma ser classificada em dois grupos: similaridade lexical e semântica. A primeira mede a similaridade das entidades através das palavras que as constituem. Nessa abordagem, normalmente são usadas soluções que medem a similaridade entre cadeias de caracteres, ou ainda heurísticas. A similaridade semântica, também tratada como similaridade semântico-estrutural, em vista da distribuição espacial da estrutura ontológica, compara as entidades de acordo com a posição das mesmas na estrutura hierárquica, buscando as relações semânticas existentes entre elas. Também se encontram, na literatura, trabalhos mais recentes com novas formas de abordar o problema, tais como os enfoques: "similaridade intrínseca" e "similaridade extrínseca" entre ontologias, com um outro ponto de vista metodológico. A similaridade intrínseca refere-se a características inerentes às entidades das ontologias e a similaridade extrínseca refere-se às relações existentes entre entidades [Heß 2006].

É importante destacar que, no contexto do presente, consideramos ontologia como uma "estrutura ontológica", ou seja, um conjunto de termos previamente definidos, associados de forma explícita por meio de relações semânticas, dispostos de forma hierárquica.

Até o momento, ao que se tem notícia, além do que será apresentado neste artigo, apenas um trabalho [Freitas 2007] foi desenvolvido para a língua portuguesa envolvendo similaridade semântica entre estruturas ontológicas, tendo por objetivo mapear termos entre estruturas ontológicas distintas e implementando uma medida denominada *SiSe* (Similaridade Semântica). Essa medida levou a coeficientes de similaridade considerados relativamente satisfatórios para o mapeamento semântico-estrutural entre estruturas ontológicas, porém não atingiu seus objetivos em alguns casos para os quais foi testada. Incluem-se aqui os resultados considerados como falsos positivos e os casos de similaridade que a medida não detectou (falsos negativos).

Partindo da *SiSe*, o presente trabalho estudou alternativas de aprimoramento, incorporando novas estratégias para o cálculo da similaridade estrutural. O foco foi fazer uso da similaridade extrínseca entre estruturas ontológicas. Quanto aos resultados, houve uma única estratégia que corrigiu todos os falsos positivos e quanto aos falsos negativos, as estratégias não apresentaram resultados muito significativos, devido à falta de uma base de dados lexicais.

Este artigo está organizado em 6 seções. A Seção 1 apresenta os conceitos sobre similaridade entre ontologias e estruturas ontológicas. A Seção 2 descreve o trabalho base e os correlatos. Na Seção 3 apresentamos um embasamento para introduzir o problema. A Seção 4 descreve as estratégias propostas. A Seção 5 apresenta os resultados e análise e a Seção 6, as conclusões sobre o trabalho.

2. *SiSe* e correlatos

2.1. Trabalho base - Medida *SiSe*

A Medida de Similaridade Semântica (*SiSe*) proposta por Freitas [Freitas 2007] adapta a proposta denominada Mapeamento Taxonômico (MT) de Maedche e Staab [Maedche and Staab 2002]. O MT faz uma comparação da similaridade entre termos de estruturas ontológicas distintas através da análise da hierarquia em que os mesmos se inserem. Desta forma, o coeficiente resultante é a similaridade semântico-estrutural

entre os termos das estruturas ontológicas. Há duas formas de calcular a similaridade na abordagem MT: *Semantic Cotopy* (*SC*) e *Common Semantic Cotopy* (*CSC*). O *SC* forma um conjunto das relações hierárquicas dos superconceitos e subconceitos do conceito em questão, independente de estes estarem inseridos na hierarquia da outra estrutura ontológica. Já o *CSC* resulta em um conjunto dos superconceitos e subconceitos comuns em ambas as hierarquias examinadas, descartando um conceito que ocorre em apenas uma hierarquia e não ocorre na outra. Com base no *SC* e *CSC*, Freitas [Freitas 2007] adaptou-os utilizando um algoritmo de *stemming* com o intuito de uniformizar os termos (por exemplo, os termos `eleições` e `eleição` possuem o *stem* (radical) `ele`). Assim, os termos serão considerados idênticos. A adaptação resultou em *SC*' e *CSC*'.

Freitas [Freitas 2007] implementou uma ferramenta para a realização dos cálculos. A ferramenta adota um roteiro que envolve a análise das linguagens utilizadas na descrição das estruturas ontológicas de entrada, abstraindo as sintaxes e normalizando-as em *XML*[1] com a representação das relações hierárquicas dos superconceitos e subconceitos. Para os testes, foram utilizados dois vocabulários: Vocabulário Controlado Básico do Senado Federal (VCBS) e o Vocabulário Controlado da USP (VCUSP). A ferramenta também permite que o usuário selecione uma das medidas desejadas e especifique um limiar para os resultados. As medidas de similaridade oferecidas são: Mapeamento Taxonômico utilizando *SC* e *CSC*, e *SiSe* utilizando *SC*' e *CSC*'. Escolhemos a medida *SiSe* utilizando *CSC*', pois apresentou melhores resultados relatados por Freitas e que foram utilizados para efeito de comparação com as estratégias apresentadas neste artigo. A *SiSe* emprega conjuntos, e é dada pela divisão da cardinalidade resultante das operações de interseção e união (vide fórmula (1)).

$$SiSe(c_1, EO_1, c_2, EO_2) = \frac{|CSC'(c_1, EO_1, EO_2) \cap CSC'(c_2, EO_2, EO_1)|}{|CSC'(c_1, EO_1, EO_2) \cup CSC'(c_2, EO_2, EO_1)|} \in [0, 1] \quad (1)$$

Onde EO_1 e EO_2 referem-se respectivamente à primeira e à segunda estrutura ontológica, e c_1 e c_2 representam um conceito da primeira estrutura e um conceito da segunda estrutura ontológica, respectivamente.

2.2. Trabalhos correlatos

Egenhofer e Rodríguez [Egenhofer and Rodríguez 2003] propõem um *framework* que possibilita a comparação de termos de uma mesma ontologia ou de ontologias distintas. A similaridade é calculada sobre a soma dos pesos das similaridades entre conjuntos de sinônimos (*synsets*), características e vizinhança. A medida inclui uma função de profundidade dos termos que corresponde à distância do termo em relação à raiz da estrutura.

[Brank et al. 2007] descreve a similaridade baseada em ancestrais comuns. Para os termos comparados, levam-se em conta os seus ancestrais, também chamados de superconceitos. A medida é uma adaptação do coeficiente de *Jaccard* que leva em consideração os superconceitos dos termos analisados.

[Isaac et al. 2007] mostra medidas estatísticas de co-ocorrência que têm por objetivo o mapeamento entre ontologias analisando as instâncias. Os autores observam que a

[1] *Extensible Markup Language*

medida de *Jaccard* possui alguns problemas quando se trata de instâncias, pois o mapeamento não distingue se os conceitos têm uma ou diversas instâncias em comum. Para corrigir isso, os autores definem a medida "*Jaccard* Corrigida" em que o objetivo é atribuir um valor menor para as anotações de co-ocorrências menos frequentes.

Felicíssimo [Felicíssimo 2004] apresenta uma estratégia, denominada CATO (Componente para Alinhamento Taxonômico entre Ontologias), que alinha automaticamente as taxonomias das ontologias de entrada. A estratégia usa como entrada duas ontologias e produz uma única ontologia como saída, onde representa as duas ontologias originais alinhadas. Faz uso de comparação lexical entre os conceitos das ontologias de entrada com sinônimos e mecanismo de poda estrutural dos conceitos e comparação entre eles.

3. Contexto de aprimoramento da medida *SiSe*

A medida *SiSe* levou a coeficientes de similaridade considerados relativamente satisfatórios para o mapeamento semântico-estrutural entre estruturas ontológicas, porém não atingiu seus objetivos em alguns casos para os quais foi testada. Incluem-se aqui os resultados inferiores a `0.7` (limiar adotado em [Freitas 2007] e também neste trabalho) e que, por um *Golden Mapping*[2] (*GM*) foram considerados similares. Estes foram aqui considerados como "falsos negativos" e os resultados superiores a `0.7`, porém ausentes do *GM*, considerados como "falsos positivos". A partir desse ponto, propusemos algumas estratégias que visam o aprimoramento da *SiSe* no que se refere aos casos citados anteriormente. Para isso, utilizamos assim como em [Freitas 2007], os 5 pares de estruturas ontológicas, provenientes de fragmentos de dois vocabulários controlados da USP (VCUSP) e Senado Federal (VCBS) do domínio do direito.

Uma primeira e importante constatação é trazida, referente à *SiSe*: para termos lexicalmente idênticos, ou seja, que possuam o mesmo *stem*, a medida estabeleceu 100% de similaridade para conjuntos que tinham em comum os seus superconceitos ou subconceitos. No entanto, casos como `direito econômico internacional` da EO_1 e `direito econômico internacional` da EO_2 merecem destaque, pois possuem superconceitos distintos (vide Figura 1).

O primeiro termo possui o conjunto (`direit`, `direitInternacPriv`, `direitEconomicInternac`) e o segundo termo o conjunto (`direit`, `direitInternacPublic`, `direitEconomicInternac`). O conjunto interseção é formado por (`direit`, `direitEconomicInternac`) e o conjunto união por (`direit`, `direitInternacPriv`, `direitEconomicInternac`, `direitInternacPublic`). O conjunto interseção possui 2 termos e o união, 4. Ao calcularmos o quociente dos dois conjuntos, chegamos a `0.5`, mas o *GM* aponta esses termos como semanticamente similares. Para tanto, seu coeficiente de similaridade

[2]A avaliação da medida *SiSe* foi realizada através de um "*Golden Mapping*", um modelo construído a partir da análise por avaliadores humanos. Para a construção desse modelo, 5 pares de estruturas ontológicas foram apresentados a 3 especialistas que indicaram os termos considerados similares. Os pares são provenientes de fragmentos de dois vocabulários controlados da USP (VCUSP) e Senado Federal (VCBS) do domínio do direito. Cada par refere-se à hierarquia de um determinado termo (por exemplo, a hierarquia do termo `direito eleitoral`). Esse modelo foi empregado em [Freitas 2007] e adotado aqui para avaliarmos as estratégias propostas.

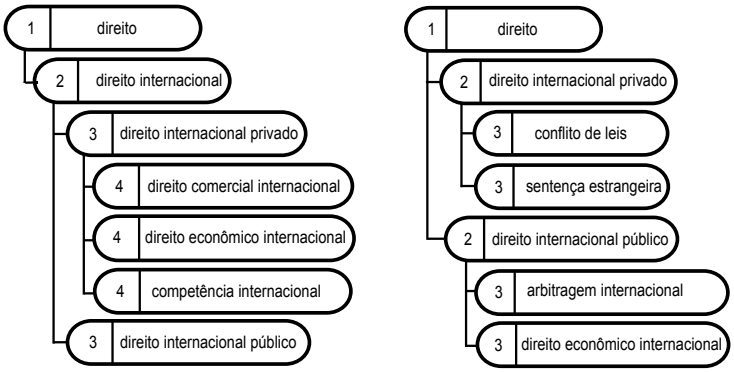

Figura 1. Extratos da EO_1 e da EO_2 (par 5)

deveria ser superior ao limiar estipulado que é `0.7`. Consideramos este caso como "falso negativo".

Uma segunda constatação refere-se aos "falsos negativos" apresentados em termos lexicalmente diferentes. Temos como exemplo: `direito internacional` da EO_1 e `direito internacional privado` da EO_2 (vide Figura 1). Estes termos são considerados similares para o *GM*, mas a medida *SiSe* não os detectou, devido a serem lexicalmente diferentes. O primeiro termo possui o conjunto (`direit, direitInternac, direitInternacPriv, direitEconomicInternac, direitInternacPublic`) e o segundo termo o conjunto (`direit, direitInternacPriv`). O conjunto interseção é formado por (`direit, direitInternacPriv`) e o conjunto união (`direit, direitInternac, direitInternacPriv, direitEconomicInternac, direitInternacPublic`). O conjunto interseção possui 2 termos e o união, 5. Ao calcularmos o quociente, chegamos a `0.4`.

Uma terceira constatação refere-se aos "falsos positivos", ou seja, aos termos que apresentam similaridade igual ou superior a `0.7` e que não foram considerados similares pelo *GM*. Trazemos aqui, o exemplo: `direito` da EO_1 e `direito eleitoral` da EO_2 (vide Figura 2).

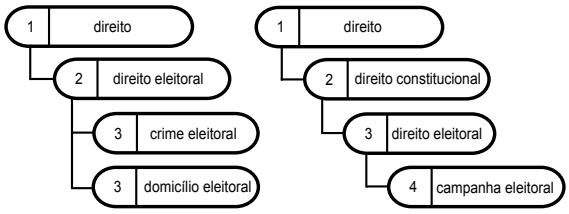

Figura 2. Extratos da EO_1 e da EO_2 (par 4)

O primeiro e segundo termo possuem o conjunto (`direit, direitEleitor`), pois estes são comuns a ambas as estruturas ontológicas, sendo que os conjuntos interseção e união são formados pelos mesmos termos. Como os conjuntos possuem 2 termos, ao calcularmos o quociente, chegamos a `1.0` (100% de similaridade).

4. As estratégias de aprimoramento propostas

A fim de corrigir os coeficientes de similaridade, são apresentadas nesta Seção as estratégias propostas, baseadas na similaridade estrutural dos termos analisados.

4.1. Fatores de adequação

Apresentamos aqui duas estratégias que visam atribuir à *SiSe* um "fator de adequação", cujo objetivo é diminuir o coeficiente de similaridade entre os termos classificados como falsos positivos, ou seja, a termos lexicalmente diferentes que foram considerados similares e que não foram apresentados no *GM* . Quanto aos termos lexicalmente idênticos, o fator considera apenas o valor da *SiSe*.

Na primeira estratégia (*FA1*), para termos lexicalmente diferentes, o fator considera o valor da *SiSe* multiplicado pelo resultado obtido com a fórmula (2). Para os termos lexicalmente idênticos, o fator de adequação será `1`.

O fator de adequação para os termos lexicalmente diferentes é representado por:

$$fator1_{t_i t_j} = \frac{\frac{1}{l_{t_i}} + \frac{1}{l_{t_j}}}{2} \tag{2}$$

Onde l_{t_i} é o nível[3] do termo t_i da EO_1 e l_{t_j} é o nível do termo t_j da EO_2.

Para a segunda estratégia (*FA2*), o fator de adequação é representado pela adaptação da fórmula utilizada em [Egenhofer and Rodríguez 2003]. Aplicamos esse novo "fator de adequação" à *SiSe*.

O fator de adequação para os termos lexicalmente diferentes é representado por:

$$fator2_{t_i t_j} = \begin{cases} \frac{nivel(t_i)}{nivel(t_i)+nivel(t_j)} & \text{se } nivel(t_i) \leq nivel\ (t_j) \\ 1 - \frac{nivel(t_i)}{nivel(t_i)+nivel(t_j)} & \text{se } nivel(t_i) > nivel(t_j) \end{cases} \tag{3}$$

Onde a primeira condição será utilizada se o nível $nivel(t_i)$ for menor ou igual ao $nivel(t_j)$, ou seja, se o nível do termo da EO_1 for menor ou igual ao nível do termo da EO_2; e a segunda condição, se $nivel(t_i)$ for maior que $nivel(t_j)$, ou seja, se o nível do termo da EO_1 for maior que o nível do termo da EO_2.

O procedimento para o cálculo da estratégia é idêntico ao anterior, ou seja, aplicamos o *fator2* à *SiSe* obtendo o resultado para *FA2*. Para os termos lexicalmente idênticos, o *fator2* terá o valor `1`.

4.2. Estratégias com busca de superconceitos e subconceitos

Sabemos que, para a formação dos conjuntos da *SiSe*, a medida leva em consideração o termo e os superconceitos e/ou subconceitos em comum a ambas as estruturas ontológicas. Com base nisso, apresentamos as estratégias *SP* e *SB* que visam melhorar os coeficientes de termos que deveriam alcançar o limiar de similaridade (falsos negativos) e corrigir os falsos positivos. Pretendemos aqui, analisar o comportamento destas

[3]O nível é constituído pela altura do termo em relação à sua raiz. Convencionamos iniciar com o nível 1 para a raiz (vide Figura 2, onde à esquerda de cada termo é representado o seu nível.

medidas de acordo com suas características e verificar se as mesmas obtiveram vantagens em relação à *SiSe*. Estas estratégias servem como uma "correção" da *SiSe*.

[Brank et al. 2007] leva em consideração os superconceitos dos termos. Por essa razão, a *SP* é apresentada como em (4).

$$SP = \frac{|CSC_{SP}(c_1, EO_1, EO_2) \cap CSC_{SP}(c_2, EO_2, EO_1)|}{|CSC_{SP}(c_1, EO_1, EO_2) \cup CSC_{SP}(c_2, EO_2, EO_1)|} \in [0, 1] \quad (4)$$

O conjunto CSC_{SP} de um termo é formado com base nos seus superconceitos que são comuns a ambas as estruturas ontológicas e o próprio termo analisado.

Seguindo o mesmo raciocínio da estratégia anterior, apresentamos a estratégia *SB* que leva em consideração os subconceitos dos termos.

$$SB = \frac{|CSC_{SB}(c_1, EO_1, EO_2) \cap CSC_{SB}(c_2, EO_2, EO_1)|}{|CSC_{SB}(c_1, EO_1, EO_2) \cup CSC_{SB}(c_2, EO_2, EO_1)|} \in [0, 1] \quad (5)$$

O conjunto CSC_{SB} de um termo é formado com base nos seus subconceitos que são comuns a ambas as estruturas ontológicas e o próprio termo analisado.

4.3. Estratégia *Jaccard* Corrigida - *JC*

Esta estratégia visa atribuir um menor coeficiente de similaridade para os termos das estruturas ontológicas distintas, e seu objetivo é corrigir o coeficiente de similaridade dos termos classificados como falsos positivos. Em [Isaac et al. 2007], uma "correção" da medida de *Jaccard* é apresentada, com a inclusão de um ajuste (valor arbitrário), o mesmo que utilizamos.

$$JC(A, B) = \frac{\sqrt{|A \cap B| \times (|A \cap B| - 0.8)}}{|A \cup B|} \quad (6)$$

Onde `A` representa o conjunto dos subconceitos e superconceitos do termo da primeira estrutura ontológica e `B` representa o conjunto dos subconceitos e superconceitos do termo da segunda estrutura ontológica.

5. Resultados e avaliação das estratégias empregadas

Os resultados obtidos pelo emprego das estratégias propostas são mostrados nas Tabelas 1 a 5, já avaliadas de acordo com o *Golden Mapping* utilizado por Freitas, fazendo-se constar precisão, abrangência e medida-F.

O par 1 possui 21 termos na EO_1 e 23 na EO_2. Ao obervarmos a Tabela 1, notamos que não houveram vantagens das estratégias em relação à *SiSe*, pois a precisão, abrangência e medida-F apresentaram os mesmos resultados.

O par 2 possui 17 termos na EO_1 e 16 na EO_2. Apenas duas estratégias apresentaram melhores resultados (vide Tabela 2). As estratégias *FA2* e *SP* resultaram em 100%

Tabela 1. Resultados para o par 1

	Termos similares	Falsos positivos	Falsos negativos	Precisão	Abrangência	Medida-F
GM	14	-	-	-	-	-
SiSe	12	-	2	100%	85.71%	92.31%
FA1	12	-	2	100%	85.71%	92.31%
FA2	12	-	2	100%	85.71%	92.31%
SP	12	-	2	100%	85.71%	92.31%
SB	12	-	2	100%	85.71%	92.31%
JC	12	-	2	100%	85.71%	92.31%

Tabela 2. Resultados para o par 2

	Termos similares	Falsos positivos	Falsos negativos	Precisão	Abrangência	Medida-F
GM	11	-	-	-	-	
SiSe	8	2	3	80%	72.73%	76.19%
FA1	8	2	3	80%	72.73%	76.19%
FA2	8	-	3	100%	72.73%	84.21%
SP	8	-	3	100%	72.73%	84.21%
SB	8	2	3	80%	72.73%	76.19%
JC	8	2	3	80%	72.73%	76.19%

de precisão, pois a ocorrência de falsos positivos foi corrigida em relação à *SiSe*. Quanto às demais estratégias, os resultados foram os mesmos.

O par 3 possui 20 termos na EO_1 e 16 na EO_2. Assim como o par 2, o par 3 obteve melhores resultados para as estratégias *FA2* e *SP* (vide Tabela 3).

Tabela 3. Resultados - par 3

	Termos similares	Falsos positivos	Falsos negativos	Precisão	Abrangência	Medida-F
GM	13	-	-	-	-	-
SiSe	9	2	4	81.82%	69.23%	75%
FA1	9	2	4	81.82%	69.23%	75%
FA2	9	-	4	100%	69.23%	81.82%
SP	9	-	4	100%	69.23%	81.82%
SB	9	2	4	81.82%	69.23%	75%
JC	9	2	4	81.82%	69.23%	75%

O par 4, que possui 40 termos na EO_1 e 34 na EO_2, em apenas um caso alcançou 100% de precisão, a *FA2* (vide Tabela 4). No entanto, a *SP* apresentou os piores resultados, pois o número de falsos positivos foi bastante expressivo. Notamos que este par não apresentou falsos negativos (todos os termos são lexicalmente idênticos). Quanto aos outros pares, todos apresentaram este tipo de ocorrência.

O par 5 que possui 28 termos na EO_1 e 19 na EO_2 apresentou o menor número de termos similares em relação ao *GM* (vide Tabela 5). O número de falsos negativos foi bastante expressivo, devido a termos serem lexicalmente diferentes e alguns casos, por pertencerem a superconceitos distintos. A única estratégia que apresentou um me-

Tabela 4. Resultados para o par 4

	Termos similares	Falsos positivos	Falsos negativos	Precisão	Abrangência	Medida-F
GM	21	-	-	-	-	-
SiSe	21	26	-	44.68%	100%	61.76%
FA1	21	2	-	91.30%	100%	95.45%
FA2	21	-	-	100%	100%	100%
SP	21	64	-	24.71%	100%	39.63%
SB	21	4	-	84%	100%	91.30%
JC	21	5	-	80.77%	100%	89.36%

lhor resultado para estes casos foi a *SB*. Quanto aos falsos positivos, as estratégias que corrigiram estas ocorrências foram: *FA1*, *FA2* e *SP*.

Tabela 5. Resultados para o par 5

	Termos similares	Falsos positivos	Falsos negativos	Precisão	Abrangência	Medida-F
GM	15	-	-	-	-	-
SiSe	7	2	8	77.78%	46.67%	58.33%
FA1	6	-	9	100%	40%	57.14%
FA2	6	-	9	100%	40%	57.14%
SP	7	-	8	100%	46.67%	63.64%
SB	8	2	7	80%	53.33%	64%
JC	7	2	8	77.78%	46.67%	58.33%

Notamos que, pela avaliação de cada par, as melhores estratégias foram a *FA2* e *SB*, para os falsos positivos e falsos negativos, respectivamente. Ao calcularmos a média ponderada para todos os pares (em conjunto) confirmamos esta afirmação (vide Tabela 6).

Tabela 6. Média ponderada para todos os pares de estruturas ontológicas

	SiSe	*FA1*	*FA2*	*SP*	*SB*	*JC*
Precisão	64.04 %	90.32 %	100 %	47.11 %	85.29 %	83.82 %
Abrangência	77.03 %	75.68 %	75.68 %	77.03 %	78.38 %	77.03 %
Medida-F	69.94 %	82.35 %	86.15 %	58.46 %	81.69 %	80.28 %

6. Conclusões

Diante dos estudos realizados, apresentamos como contribuição deste trabalho, a criação e adaptação de estratégias para melhoria nos resultados do cálculo da similaridade estrutural. Ao aplicarmos as estratégias observamos que duas apresentaram excelentes resultados, *FA2* e *SB*, sendo que podemos substituir a *SiSe* pela *FA2* e as demais estratégias poderão ser utilizadas pela aplicação de heurísticas, ou seja, aplicar as estratégias ou combiná-las de acordo com as características dos termos (por exemplo, termos lexicalmente diferentes ou idênticos). Com isso, uma base de dados lexicais, como uma lista de sinônimos poderá ser utilizada para encontrar um maior número de correspondências entre os termos lexicalmente diferentes. Atualmente, existem trabalhos em andamento como o TEP2[4], cujos dados ainda não são suficientes para auxiliar no cálculo da simi-

[4]Base de dados lexicais da língua portuguesa que está sendo desenvolvida pelo NILC-USP.

laridade intrínseca. O uso do Vocabulário Controlado do Senado Federal (VCBS) para recuperar sinônimos também não se mostrou útil. Cabe salientar que nos deparamos com uma situação interessante, tão importante quanto a escassez de ontologias para a língua portuguesa foi a escassez de léxicos para esse idioma.

Apesar de focarmos na língua portuguesa, este trabalho é aplicável a outros idiomas.

Referências

Brank, J., Grobelnik, M., and Mladenic, D. (2007). *"Automatic Evaluation of Ontologies"*, chapter 11, pages 193–219. Springer London.

Egenhofer, M. J. and Rodríguez, M. A. (2003). "Determining Semantic Similarity among Entity Classes from Different Ontologies". *IEEE Transations on Knowledge and Data Engineering*, 15(2):442–456.

Ehrig, M. (2007). *"Ontology Alignment: Bridging the Semantic gap"*. Semantic Web and Beyond: Computing for Human Experience.Springer-Verlag New York, Inc., New York, NY, USA.

Felicíssimo, C. H. (2004). "Interoperabilidade Semântica na Web: Uma Estratégia para o Alinhamento Taxonômico de Ontologias". 180f. Dissertação (Mestrado em Informática), PUC-Rio, Rio de Janeiro, 2004.

Freitas, J. B. (2007). "SiSe: Medida de Similaridade Semântica entre Ontologias em Português". 93f. Dissertação (Mestrado em Ciência da Computação) - Faculdade de Informática, PUCRS, Porto Alegre, 2007.

Heß, A. (2006). "An Iterative Algorithm for Ontology Mapping Capable of Using Training Data". In *ESWC'06: Proceedings of the 3rd European Semantic Web Conference*, pages 19–33.ACM.

Isaac, A., der Meij, L. V., Schlobach, S., and Wang, S. (2007). "An Empirical Study of Instance-based Ontology Matching". In Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L. J. B., Golbeck, J., Mika, P., Maynard, D., Schreiber, G., and Cudré-Mauroux, P., editors, *ISWC/ASWC2007: Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference, Busan, South Korea*, volume 4825 of *LNCS*, pages 252–266, Berlin, Heidelberg.Springer Verlag.

Maedche, A. and Staab, S. (2002). "Measuring Similarity between Ontologies". In *EKAW '02: Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management . Ontologies and the Semantic Web*, volume 2473, pages 251–263.Springer-Verlag.