3
Alinhamento de tesauros

3.1
Introdução

Um catálogo é um banco de dados que armazena informações sobre um conjunto de objetos classificados usando termos de um esquema de classificação como, por exemplo, um tesauro. Portanto, o projeto de um mediador para uma coleção de catálogos requer o alinhamento de tesauros. Da mesma forma, quando as instâncias de um catálogo precisam ser carregadas para outro catálogo, a etapa de transformação de dados deve incorporar o alinhamento *a priori* dos tesauros, pois as instâncias do catálogo de origem precisam ser reclassificadas utilizando o esquema de classificação do catálogo de destino.

Neste capítulo são apresentadas abordagens para alinhamento de tesauros utilizando instâncias equivalentes como evidências. É importante ressaltar que estas abordagens assumem que é possível reconhecer quando instâncias de diferentes catálogos representam um mesmo objeto. Para detectar instâncias equivalentes é necessário utilizar o valor de um atributo, ou de um conjunto de atributos, que sirva como identificador único dos objetos. Em aplicações de comércio eletrônico, por exemplo, pode ser usado o código de produto (*part number*), se ambos os catálogos armazenarem tal informação. De forma similar, em aplicações que armazenam dados geográficos, a localização espacial dos objetos pode ser usada como identificador único dos objetos.

Note que, nestas abordagens, assumimos que a heterogeneidade dos esquemas conceituais das fontes utilizadas já foi resolvida, manualmente ou através de uma técnica de alinhamento de esquemas conceituais. Nesta tese, o alinhamento de esquemas conceituais é discutido no Capítulo 4.

A seção 3.2 introduz uma abordagem *a priori* para alinhamento de tesauros. Em (Brauner et al., 2007a) é apresentada uma aplicação desta abordagem. A seção 3.3 apresenta uma abordagem adaptativa para alinhamento de tesauros. Em (Brauner et al., 2006) é investigado o acesso mediado a uma coleção de catálogos utilizando esta abordagem.

3.2
Abordagem *a priori*

A abordagem, descrita nesta seção, tem por objetivo alinhar, de forma *a priori*, esquemas de classificação (tesauros) diferentes, utilizados por catálogos distintos. Para isso, são estimadas taxas de mapeamento entre os termos dos tesauros a partir de um conjunto de instâncias equivalentes coletadas dos catálogos.

A abordagem *a priori* de alinhamento de tesauros é baseada no processo ilustrado na Figura 9.

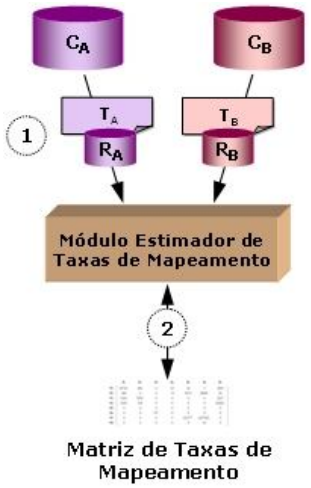

Figura 9 – O processo de alinhamento *a priori* de tesauros.

Considere dois catálogos, C_A e C_B, e assuma que eles adotem os tesauros T_A e T_B, respectivamente. O objetivo é carregar os dados de C_A para C_B. Para isso, é necessário criar mapeamentos dos termos do tesauro T_A para os termos do tesauro T_B para que as instâncias de C_A possam ser reclassificadas utilizando os termos de T_B.

Duas instâncias $c_a \in C_A$ e $c_b \in C_B$ são *equivalentes*, denotado por $c_a \equiv c_b$, quando elas representam o mesmo objeto do mundo real. O procedimento exato para computar a equivalência de instâncias depende do domínio de aplicação.

O processo coleta inicialmente evidências para os mapeamentos a partir de conjuntos de instâncias equivalentes, $R_A \subseteq C_A$ e $R_B \subseteq C_B$ (passo 1 da Figura 9).

Dado um par de instâncias equivalentes, c_a e c_b, é possível afirmar que elas computam uma evidência de que seus respectivos tipos, $t_a \in T_A$ e $t_b \in T_B$,

mapeiam. Então, deve-se computar a Matriz de Taxas de Mapeamento (passo 2 da Figura 9) utilizando a fórmula (1):

$$P(t_a, t_b) = \frac{n(t_a, t_b)}{n(t_a)} \quad (1)$$

Dados os conjuntos de instâncias equivalentes, R_A e R_B, as taxas de mapeamento são calculadas a partir dos seguintes índices:

1. *$n(t_a,t_b)$:* soma das ocorrências de pares de objetos c_a e c_b, tais que:
   - $c_a \in R_A$ e $c_b \in R_B$
   - $c_a \equiv c_b$
   - t_a e t_b são os tipos de c_a e c_b, respectivamente
2. *$n(t_a)$:* soma das ocorrências de instancias em R_A com tipo t_a.
3. *$P(t_a,t_b)$:* taxa de mapeamento do termo t_a em t_b, usando a Equação (1) para estimar a freqüência que o termo t_a alinha ao termo t_b. Esta taxa é calculada para todos os pares de termos $t_a \in T_A$ e $t_b \in T_B$ que ocorrem em R_A e R_B.

Note que o procedimento acima é simétrico. Portanto, o processo pode ser facilmente adaptado para computar a freqüência com que os termos de T_B mapeiam com os termos de T_A. Para isso, basta computar *$P(t_b,t_a)$*, utilizando *$n(t_b)$* ao invés de *$n(t_a)$*, e alterar o denominador da fórmula (1) para *$n(t_b)$*. As taxas são armazenadas em uma Matriz de Taxas de Mapeamento.

Para descobrir o limiar para as taxas de mapeamento, foi realizado um processo de validação cruzada, detalhado na seção 3.4.2. Durante este processo, obteve-se os melhores índices para o limiar 0.4. Obteve-se 95.4% de precisão nos mapeamentos de T_A para T_B, e 97% de precisão nos mapeamentos de T_B para T_A. Após o processo de validação, foi realizado um teste do modelo. Assim, pares de termos com taxas de mapeamento acima deste limiar retratam os alinhamentos corretos descobertos.

De posse das taxas de mapeamento e do limiar descoberto, é possível criar um mediador que use estas taxas para mediar consultas a fontes distintas ou, ainda, carregar dados de um catálogo para outro, para isso basta re-classificar as instâncias utilizando os termos que possuírem taxas superiores ao limiar descoberto. Nesta última, o processo de reclassificação fica a critério do usuário, isto é, as instâncias poderão manter a classificação antiga e incluir a

nova classificação como um novo atributo, ou simplesmente poderão assumir apenas a nova classificação, descartando a antiga.

3.3
Abordagem adaptativa

Esta abordagem usa um estimador adaptativo para alinhar termos de esquemas de classificação (tesauros) diferentes. A abordagem atribui, gradualmente, pesos aos relacionamentos entre os termos de tesauros distintos, através do pós-processamento dos resultados das consultas submetidas a um mediador de consultas durante seções de usuário.

A abordagem adaptativa de alinhamento de tesauros é baseada no processo ilustrado na Figura 10.

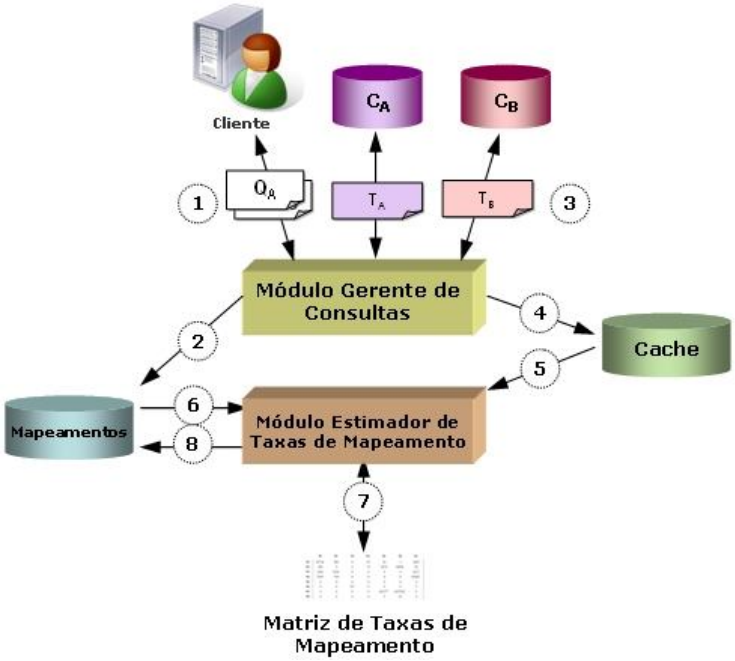

Figura 10 – O processo de alinhamento *adaptativo* de tesauros.

Por simplicidade, assume-se apenas dois catálogos, C_A e C_B, armazenando objetos num mesmo domínio, classificados usando os tesauros T_A e T_B, respectivamente. Neste contexto, deseja-se mapear os termos de T_A em termos de T_B. Entretanto, a discussão pode ser generalizada para mapeamentos bidirecionais e utilizando mais de dois catálogos.

O processo começa com uma *sessão de usuário*, composta por uma consulta Q_A sobre C_A ou Q_B sobre C_B, ou por um par de consultas Q_A sobre C_A e

Q_B sobre C_B, submetidas através do mediador (passo 1 da Figura 10). Assume-se que, em uma sessão de usuário que contenha um par de consultas, estas consultas tentam recuperar objetos de C_A e C_B que são classificados de forma similar como, por exemplo, consultas por objetos do tipo “mountain”.

Recebida a primeira consulta, o mediador verifica a existência de mapeamentos para o termo utilizado na consulta (passo 2) através do *Módulo Gerente de Consultas* (MGC). Se não existirem mapeamentos, é solicitado ao usuário que formule a segunda consulta, utilizando termos do tesauro da segunda fonte. Se existirem mapeamentos, o MGC poderá recomendar o termo a ser utilizado para a segunda consulta e formulá-la automaticamente.

Feito isso, o MGC submete as consultas às fontes e retorna o resultado das consultas ao usuário (passo 3).

O resultado das consultas é armazenado em um cache local (passo 4).

A seguir, o mediador computa as taxas de mapeamento com base nas instâncias equivalentes recuperadas dos resultados armazenados em cache (passo 5). Isto é feito através do *Módulo Estimador de Taxas de Mapeamento* (METM). Recorde que duas instâncias $c_a \in C_A$ e $c_b \in C_B$ são *equivalentes*, denotado por $c_a \equiv c_b$, quando elas representam o mesmo objeto do mundo real. O procedimento exato para computar a equivalência de instâncias depende do domínio de aplicação.

Se existirem mapeamentos anteriormente calculados, identificados no passo 2, o METM recupera as evidências anteriormente computadas (passo 6) que serão utilizadas para computar a *taxa de mapeamento*.

Dados dois termos $t_a \in T_A$ e $t_b \in T_B$, o mediador computa a *matriz de taxas de mapeamento* (passo 7) utilizando a fórmula (2), que estima a freqüência com que o termo t_a mapeia com o termo t_b.

Para finalizar, os novos valores das taxas de mapeamento, bem como os valores das evidências, são atualizados no banco de dados de mapeamentos (passo 8).

$$P(t_a, t_b) = \frac{\Delta(t_a, t_b) + \alpha\left(n(t_a, t_b) + \psi\right)}{\Delta(t_a) + \alpha\left(n(t_a) + 1\right)} \quad (2)$$

Onde:

- α coeficiente que assume um dos valores do conjunto {0.01, 0.1, 0, 1, 10, 100}. Este coeficiente é calibrado durante o processo de

validação (seção 3.4.2). Neste contexto, 100 indica dar maior importância às evidências antigas enquanto 0.01 indica dar maior importância às novas evidências.

$\psi = \frac{1}{|T_B|}$ coeficiente de suavização dado pelo inverso do tamanho do tesauro do segundo termo. Obs: o tamanho do tesauro é dado pelo número de termos que ele possui.

Dados dois resultados de consultas, Q_A e Q_B, de uma mesma sessão de usuário recuperados do cache, o mediador computa os conjuntos de instâncias equivalentes R_A e R_B. A partir destes conjuntos, os seguintes índices são calculados e armazenados:

1. $n(t_a,t_b)$*:* soma das ocorrências de pares de objetos c_a e c_b, tais que:
   - $c_a \in R_A$ e $c_b \in R_B$, obtidos de uma mesma seção de usuário
   - $c_a \equiv c_b$
   - t_a e t_b são os tipos de c_a e c_b, respectivamente
2. $n(t_a)$*:* soma das ocorrências de $c_a \in R_A$ onde t_a é o tipo de c_a.

Note que, na fórmula (2), os valores para $\Delta(t_a,t_b)$ e $\Delta(t_a)$ são calculados como $n(t_a,t_b)$ e $n(t_a)$, representando as evidências computadas a partir dos resultados de uma consulta atual, enquanto os valores de $n(t_a,t_b)$ e $n(t_a)$ na fórmula indicam os valores computados das ocorrências anteriores de t_a e t_b. Note ainda que, na primeira ocorrência de um par t_a e t_b, os valores de $n(t_a,t_b)$ e $n(t_a)$ serão nulos, pois não existiam evidências anteriores para este par de termos. Por isso, o mediador precisa armazenar estes valores em um banco de dados local (passo 8 da Figura 10).

Assim, $P(t_a,t_b)$, a taxa de mapeamento do termo t_a em t_b, é calculada usando a fórmula (2), um estimador para a freqüência com que o termo t_a alinha com o termo t_b. Esta taxa é calculada para todos os pares de termos $t_a \in T_A$ e $t_b \in T_B$ que ocorrem em R_A e R_B.

Note que o procedimento acima é simétrico. Portanto, o processo pode ser facilmente adaptado para computar a freqüência com que os termos de T_B mapeiam nos termos de T_A. Para isso, basta computar $\Delta(t_b)$ e $n(t_b)$, ao invés de $\Delta(t_a)$ e $n(t_a)$, e alterar o denominador da fórmula (2) para $\Delta(t_b)$ e $n(t_b)$.

Para descobrir o limiar para as taxas de mapeamento, foi realizado um processo de validação cruzada, detalhado na seção 3.4.3. Durante este processo, foi obtida uma precisão de 95.4% e um limiar de 0.4. Assim, pares de

termos com taxas de mapeamento acima deste limiar retratam alinhamentos postulados como corretos.

O processo, ilustrado pela Figura 10, é aplicado a um mediador de consultas a fontes de dados com tesauros heterogêneos. A Figura 11 mostra a arquitetura proposta para um mediador utilizando esta abordagem.

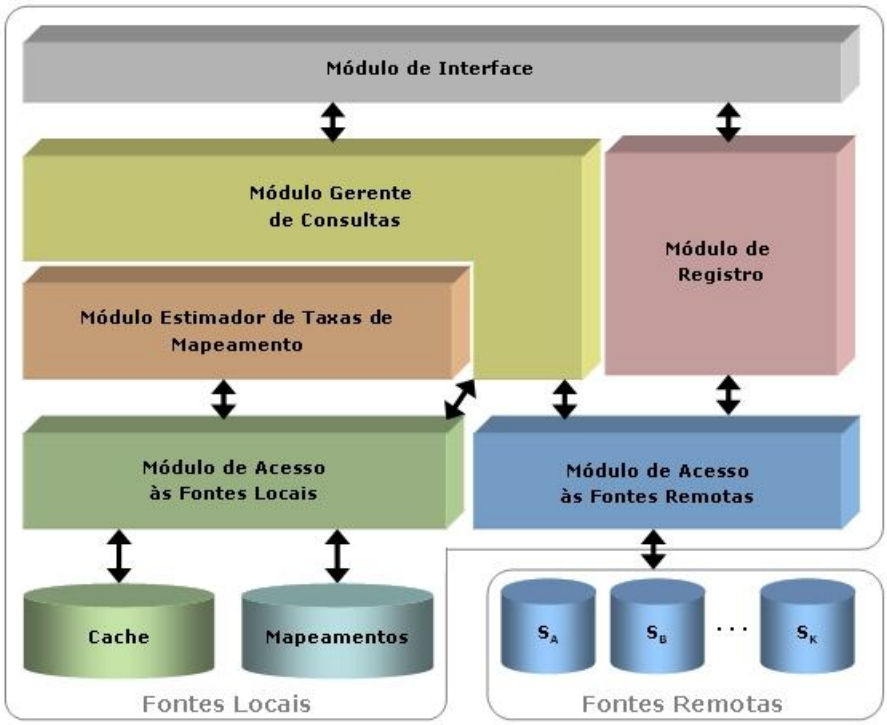

Figura 11 – Arquitetura proposta para um mediador utilizando a abordagem *adaptativa* para alinhamento de tesauros.

O *Módulo de Interface* (MI) é responsável pela comunicação entre os usuários e o mediador. O MI recebe as consultas dos usuários e retorna seus resultados. Ele se comunica com o *Módulo de Registro* (MR) para cadastrar, editar e excluir fontes de dados. No momento do cadastramento, o usuário já deve especificar os mapeamentos dos elementos dos esquemas conceituais, o(s) atributo(s) identificador(es) e o esquema de classificação. O MI se comunica com o *Módulo Gerente de Consultas* (MGC). O MGC é responsável por decompor as consultas dos usuários em subconsultas, reescrevendo-as no formato padrão das fontes de dados cadastradas, e submetê-las. O MGC comunica-se com a camada de *Módulos de Acesso* (adaptadores) para acessar as fontes de dados locais e remotas. Durante o processo de decomposição das consultas, o MGC comunica-se com o banco de dados de mapeamentos

(através do *Módulo de Acesso às Fontes Locais* - MAFL) para recuperar os mapeamentos existentes e formular as consultas no formato das fontes de dados. Além disso, o MGC comunica-se com o MAFL para armazenar as respostas das consultas num cache local do mediador. O *Módulo Estimador de Taxas de Mapeamento* (METM) é um módulo autônomo, responsável por acessar o cache para computar as taxas de mapeamento para os termos dos esquemas de classificação.

Gazola et al. (2007) apresenta uma implementação da arquitetura proposta para um mediador de catálogos de dados geográficos utilizando a abordagem adaptativa para alinhamento de tesauros.

O mediador implementado oferece suporte a dois tipos de consulta: por classificação ou por palavra-chave. Na consulta por classificação, o usuário seleciona um termo *t* do tesauro de sua preferência, dentre os tesauros disponíveis das fontes cadastradas. O mediador usa os mapeamentos já descobertos, se existentes, para mapear *t* em termos dos outros tesauros. Se ainda não existirem mapeamentos, o mediador solicita ao usuário para informar qual o termo correspondente do outro tesauro. A consulta é então processada e os resultados armazenados em cache são analisados, gerando novas taxas de mapeamento que serão gradualmente atualizadas a cada nova consulta. Numa consulta por palavra-chave, o usuário entra com um nome de lugar, que será consultado pelo atributo *nome* de cada catálogo de dados geográficos. Os resultados são analisados da mesma forma da consulta por classificação. Para detectar as instâncias equivalentes o mediador implementado utiliza uma comparação sintática simples dos atributos de latitude e longitude, presentes em ambas as fontes utilizadas.

3.4
Validação e testes

3.4.1
Introdução

Nesta seção são exemplificados casos de heterogeneidade entre dois tesauros reais. Os tesauros apresentados aqui são utilizados nos experimentos para validação e teste apresentados ao longo desta seção.

Um *gazetteer* é um catálogo de nomes geográficos contendo as representações espaciais dos objetos e outras informações descritivas (Hill et al., 1999). Geralmente, os gazetters possuem seus objetos classificados com

base em um tesauro de feições geográficas. Nesta tese, utilizamos dois *gazetteers* disponíveis na Web para avaliar as abordagens através de experimentos práticos com dados reais: o *GEOnet Names Server* e o *Alexandria Digital Library Gazetteer*.

O *GEOnet Names Server* (GEOnet) (GNS, 2008) é um *gazetteer* que fornece acesso ao banco de dados oficial de nomes estrangeiros utilizados pelo governo dos Estados Unidos da América. O GEOnet é mantido pela Agência Nacional de Inteligência Geoespacial (*National Geospatial-Intelligence Agency* - NGA) e seus dados são aprovados pelo Comitê de Nomes Geográficos do governo dos Estados Unidos (*U.S. Board on Geographic Names*). O GEOnet contém em torno de 5.5 milhões de nomes a respeito de 4 milhões de objetos geográficos incluindo cidades, estados, países, rios, montanhas, etc. Os dados do GEOnet são classificados utilizando um tesauro contendo 645 termos classificados sob 9 classes principais (Tabela 1). A Tabela 2 apresenta um fragmento do tesauro GEOnet.

O *Alexandria Digital Library* (ADL) (ADL, 1999) é um catálogo de imagens, mapas e dados geográficos, incluindo um *gazetteer* com aproximadamente 5.9 milhões de nomes geográficos. O ADL Gazetteer (Hill et al., 1999) é baseado num padrão de conteúdo próprio, chamado *ADL Gazetteer Content Standard*, e seus objetos são classificados utilizando o *ADL Feature Type Thesaurus* (ADL FTT) (ADL FTT, 2002). O ADL FTT possui 1.262 termos, organizados hierarquicamente e relacionados através de uma coleção estendida de relacionamentos de tesauros. Nesta tese, foram considerados apenas os termos preferenciais, totalizando 210 termos classificados sob 6 classes principais (Tabela 1). A Tabela 3 apresenta um fragmento do ADL FTT.

Tabela 1 – Classes principais dos tesauros GEOnet e ADL FTT.

GEOnet	ADL FTT
Administrative Region	Administrative Areas
Area	Hydrographic Features
Hydrographic	Land Parcels
Hypsographic	Manmade Features
Populated Place	Physiographic Features
Spot Features	Regions
Streets / Highways / Roads	
Undersea	
Vegetation	

Tabela 2 – Fragmento do tesauro do GEOnet.

Code	Name	Designation Text	Class
ADM1	first-order administrative division	a primary administrative division of a country, such as a state in the United States	A - Administrative Region
FLLS	waterfall(s)	a perpendicular or very steep descent of the water of a stream	H – Hydrographic
FRST	forest(s)	an area dominated by tree vegetation	V – Vegetation
INDS	industrial area	an area characterized by industrial activity	L – Área
PPL	populated place	a city, town, village, or other agglomeration of buildings where people live and work	P - Populated Place
PRMN	promenade	a place for public walking, usually along a beach front	R - Streets / Highways / Roads
RDGE	ridge(s)	a long narrow elevation with steep sides, and a more or less continuous crest	T - Hypsographic
RISU	rise	a broad elevation that rises gently, and generally smoothly, from the sea floor	U – Undersea
RSTN	railroad station	a facility comprising ticket office, platforms, etc. for loading and unloading train passengers and freight	S - Spot Feature

Tabela 3 – Fragmento do tesauro ADL FTT.

Name	Designation Text	Class
bays	indentations of a coastline or shoreline enclosing a part of a body of water; bodies of water partly surrounded by land.	Hydrographic Features
islands	tracts of land smaller than a continent, surrounded by the water of an ocean, sea, lake or stream.	Regions
parks	places or areas developed for public use or recreation.	Manmade Features
populated places	places or areas with clustered or scattered buildings and a permanent human population.	Administrative Áreas
religious facilities	places built for the observance of faith and the pursuit of a religious life.	Manmade Features
ridges	elevations with a narrow, elongated crest which can be part of a hill or mountain.	Physiographic features
streams	linear bodies of water flowing on the Earth's surface.	Hydrographic Features

Ao longo desta seção, o ADL Gazetteer e o GEOnet Names Server são identificados por C_A e C_B, respectivamente, e seus tesauros por T_A e T_B.

A Tabela 4 mostra exemplos de objetos equivalentes extraídos de C_A e C_B. Note que um mesmo objeto, por exemplo “Pelotas”, é classificado como “populated places” na fonte C_A e como “PPL” na fonte C_B. Este objeto,

identificado de forma unívoca em ambas as fontes através dos atributos de latitude e longitude, computa uma evidência de que o termo “populated places” do tesauro T_A alinha com o termo “PPL” do tesauro T_B. Ainda neste exemplo, temos mais duas evidências deste alinhamento, dadas pelos objetos “Bagé” e “Macaé”, totalizando assim três evidências concretas de que os termos “populated places” e “PPL” alinham.

Tabela 4 – Exemplos de objetos equivalentes extraídos da ADL Gazetteer (C_A) e do GEOnet Names Server (C_B).

Instance name	Latitude	Longitude	T_A term	T_B term
Estado do Rio Grande do Sul	-30	-54	administrative areas	ADM1
Praia do Cassino	-32.1917	-52.1583	beaches	BCH
Praia Vermelha	-22.9667	-43.1667	beaches	BCH
Igreja Nossa Senhora da Paz	-32.145	-52.1033	religious facilities	CH
Cachoeira das Almas	-30.05	-52.8	waterfalls	FLLS
Morro da Urca	-22.95	-43.1667	mountains	HLL
Ilha das Flores	-29.9833	-51.25	islands	ISL
Corcovado	-21.8	-42.6833	mountains	MT
Bagé	-31.3333	-54.1	populated places	PPL
Pelotas	-31.7667	-52.3333	populated places	PPL
Macaé	-22.3833	-41.7833	populated places	PPL
Parque Nacional da Tijuca	-22.95	-43.2833	parks	PRK
Ponta do Retiro	-31.98	-52.0533	reference locations	PT
Serra da Taquara	-22.2	-43.8833	ridges	RDGE
Estação Leônidas Assis Brasil	-30.0333	-54.6667	railroad features	RSTN
Parada do Salso	-30.9833	-54.55	railroad features	RSTP
Barragem do Capigui	-28.35	-52.2	reservoirs	RSV
Rio São Pedro	-22.6667	-43.6333	streams	STM

Nas seções seguintes serão apresentados dois experimentos utilizando os tesauros apresentados aqui. Na seção 3.4.2 é apresentado um experimento realizado com a finalidade de carregar os dados do *gazetteer* C_B para o *gazetteer* C_A. Para isso, é utilizada a abordagem de alinhamento *a priori* introduzida na seção 3.2. A seção 3.4.3 apresenta um experimento realizado com a finalidade de mediar consultas entre os *gazetteers* C_A e C_B. Para isso, é utilizada a abordagem de alinhamento adaptativa introduzida na seção 3.3.

3.4.2
Validação e teste da abordagem *a priori*

Para definir o limiar que determina mapeamentos corretos entre termos a partir das taxas de mapeamento calculadas pela fórmula (1), foi executado um processo de validação cruzada em seis etapas e o teste do modelo (fórmula e limiar). O procedimento de validação cruzada e teste foi aplicado tanto para os mapeamentos de T_A para T_B quanto de T_B para T_A .

Para isso, foram coletados sete conjuntos de instâncias equivalentes de C_A e C_B (veja a legenda no canto superior esquerdo da Figura 12). Seis destes conjuntos foram considerados “conjuntos de afinação” (*Tuning sets*) e utilizados durante a validação cruzada para descobrir com qual limiar a fórmula obtém maior precisão; o conjunto restante foi considerado “conjunto de testes” (*Testing set*) e utilizado durante o teste, ao qual é aplicado o modelo obtido, calculando-se assim os índices de precisão e cobertura do modelo.

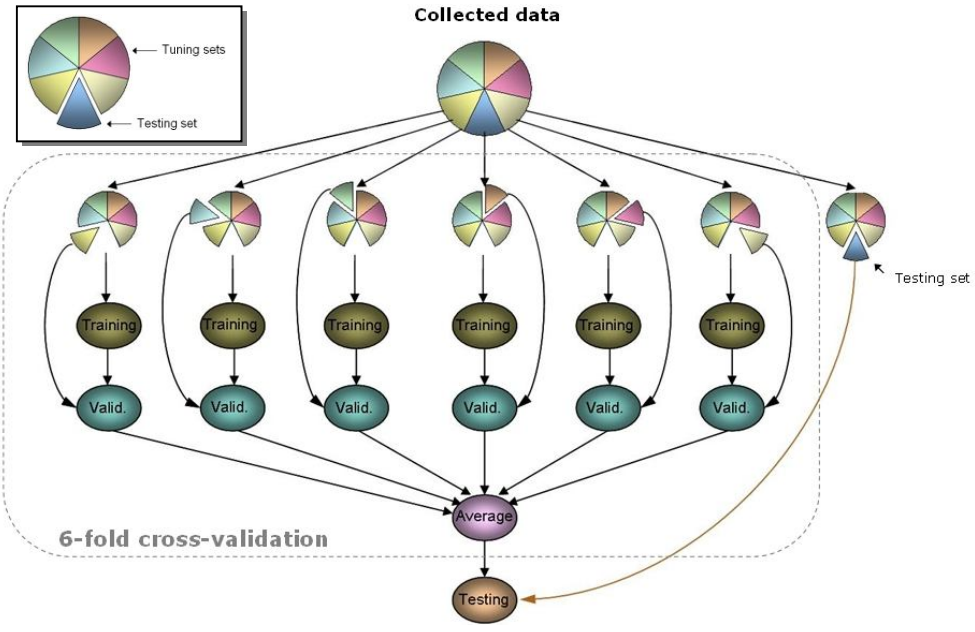

Figura 12 – Processo de validação cruzada e teste.

Na Figura 12 podemos ver o processo de validação e teste da abordagem *a priori*. A validação cruzada (etapas delimitadas pela linha tracejada na Figura 12) é aplicada em seis etapas. Em cada uma das etapas, um dos sub-conjuntos do conjunto de afinação é separado dos demais e chamado de conjunto de validação. Os cinco conjuntos restantes são unificados e o conjunto restante é

chamado de conjunto de treinamento. O conjunto de validação é etiquetado manualmente com “1” e “0” para indicar quais os pares de termos são mapeamentos corretos ou incorretos, respectivamente. A Tabela 5 mostra um exemplo de conjunto de validação. O processo de validação inicia com a fase de treinamento (*Training*), onde o modelo é aplicado ao conjunto de treinamento, gerando taxas de mapeamento para os pares evidenciados. A Tabela 6 mostra um exemplo dos resultados obtidos na fase de treinamento.

Tabela 5 – Exemplo de um conjunto de validação.

t_a	t_b	flag(t_a,t_b)	flag(t_b,t_a)
mountains	BAY	0	0
beaches	BCH	1	1
islands	ISL	1	1
physiographic features	RK	0	1
physiographic features	RKS	0	1
physiographic features	UPLD	1	1
populated places	PPL	1	1
populated places	ISL	0	0
power generation sites	PS	1	1
ridges	RDGE	1	1
waterfalls	FLLS	1	1

Tabela 6 – Exemplo de um conjunto de treinamento.

t_a	t_b	$n(t_a,t_b)$	P(t_a,t_b)	P(t_b,t_a)
agricultural sites	FRM	32	1.0000	1.0000
bays	BAY	2	1.0000	0.6667
forests	RESF	1	1.0000	1.0000
physiographic features	RK	1	1.0000	0.0833
physiographic features	RKS	1	1.0000	0.0833
physiographic features	UPLD	10	1.0000	0.8333
populated places	PPL	938	0.9621	0.9260
populated places	STM	56	0.0557	0.0553
reference locations	PT	12	0.9231	0.8000
reference locations	STM	3	0.0030	0.2000
rivers	STMA	52	0.9811	0.9811
streams	STM	936	0.9313	0.9213
waterfalls	FLLS	234	0.9710	0.9512
waterfalls	STM	8	0.0080	0.0325

O próximo passo é a fase de validação (*Validation*) onde os resultados obtidos na fase de treinamento são cruzados com os valores etiquetados do conjunto de validação. Nesta fase foram definidos limiares para computar os índices de mapeamentos acertados. Estes índices são levantados a partir dos mapeamentos propostos durante o treinamento e confrontados com os mapeamentos existentes no conjunto de validação. Os limiares utilizados pertencem ao subconjunto discreto do intervalo [0,1], com passo de tamanho 0.1, representado por L = {0.0, 0.1,..., 0.9, 1.0}. Assim, o processo de validação cruzada foi executado dez vezes, utilizando cada um elementos do conjunto L, de forma a descobrir com qual limiar o modelo obtêm maior índice de precisão.

A Figura 13 e a Figura 14 mostram os gráficos contendo os índices de precisão obtidos durante o processo de validação cruzada com mapeamentos de T_A para T_B e T_B para T_A, respectivamente. Nestes gráficos, é possível notar que se obteve os melhores índices para o limiar 0.4, tanto nos mapeamentos de T_A para T_B (Figura 13), onde se obteve 95.4% de precisão, quanto nos mapeamentos de T_B para T_A (Figura 14), onde se obteve 97% de precisão.

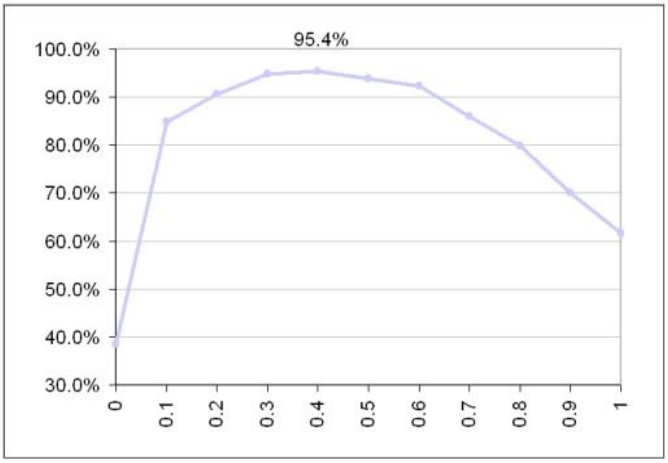

Figura 13 – Gráficos de “Precisão x Limiar” para abordagem de alinhamento *a priori* ($T_A \rightarrow T_B$).

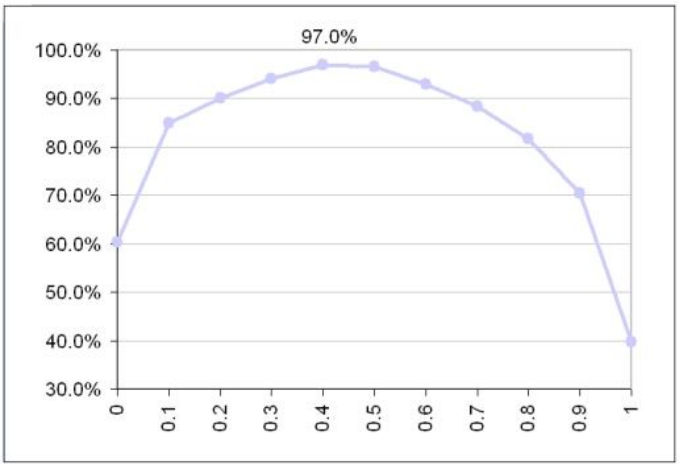

Figura 14 – Gráficos de “Precisão x Limiar” para abordagem de alinhamento *a priori* ($T_B \rightarrow T_A$).

De posse do limiar, foi executada a etapa de teste do modelo. Nesta etapa, o conjunto de testes é primeiramente etiquetado manualmente e então submetido ao modelo. Os resultados são comparados às etiquetas e, então, os índices de precisão e cobertura são calculados.

Para os testes utilizamos uma coleção de 2.920 instâncias equivalentes entre C_A e C_B, sendo 1.998 instâncias de C_A, e 2.759 instâncias de C_B, englobando objetos que representam feições geográficas de estados da região norte do Brasil (Pará, Amapá, Roraima, Amazonas, Acre e Rondônia). Para computar as instâncias equivalentes foi utilizada uma comparação sintática simples dos atributos de latitude e longitude, presentes em ambas as fontes utilizadas. A Tabela 7 mostra alguns exemplos das instâncias equivalentes coletadas.

Tabela 7 – Exemplos de objetos equivalentes extraídos da ADL Gazetteer (C_A) e do GEOnet Names Server (C_B) e suas classificações segundo T_A e T_B.

C_A Instance name	C_B instance name	Latitude	Longitude	T_A term	T_B term
N. S. de Nazare, Escola	Escola N. S. de Nazare	-4.7361	-62.1517	educational facilities	SCH
Caxiuana, Reserva Florestal de	Reserva Florestal de Caxiuana	-2.0833	-51.85	forests	RESF
Amazonia, Parque Nacional da	Parque Nacional da Amazonia	-3.8	-56.6667	parks	PRK
Porto Alegre	Porto Alegre	-8.95	-67.8333	populated places	PPL
Transamazonica, Rodovia	Rodovia Transamazonica	-5.3333	-49.1167	roadways	RD
Amazonia, Rio	Rio Amazonia	-7.2833	-61.9667	streams	STM

C_A Instance name	C_B instance name	Latitude	Longitude	T_A term	T_B term
Atua, Rio	Rio Atua	-1.5333	-49.0667	streams	STM
Bacajai, Rio	Rio Bacajai	-3.45	-51.8833	streams	STM
Gorotire, Reserva Florestal de	Reserva Florestal de Gorotire	-7.5	-52	tribal areas	RESV
Reserva Indigena Kararao	Reserva Indigena Kararao	-4.1675	-52.8861	tribal areas	RESV
Muira, Cachoeira	Cachoeira Muira	-1.7333	-54.3833	waterfalls	FLLS
Xateturu, Cachoeira	Cachoeira Xateturu	-6.5833	-52.15	waterfalls	FLLS

A partir deste conjunto de dados, foram computadas as taxas de mapeamento $P(t_a, t_b)$, de T_A para T_B, e $P(t_b, t_a)$, de T_B para T_A (passo 2 da Figura 9 da seção 3.2). A Tabela 8 mostra alguns exemplos das taxas de mapeamento obtidas.

Tabela 8 – Exemplos dos alinhamentos obtidos via abordagem *a priori*.

T_A term	T_B term	$n(t_b,t_a)$	$P(t_a,t_b)$	$P(t_b,t_a)$
agricultural sites	FRM	32	1.0000	1.0000
islands	ISL	213	0.9383	0.9861
lakes	LK	169	0.9235	1.0000
mountains	HLL	27	0.7941	1.0000
populated places	PPL	938	0.9260	0.9621
populated places	STM	56	0.0553	0.0557
rapids	RPDS	28	1.0000	0.9655
streams	CRKT	6	0.0059	1.0000
streams	PPL	33	0.0325	0.0338
streams	STM	936	0.9213	0.9313
waterfalls	FLLS	234	0.9512	0.9710

Como resultado, foram obtidos 26 pares de termos corretamente alinhados de T_A para T_B com índice de precisão de 89.7% e cobertura de 81.3% (dados os 32 pares existentes e os 29 pares propostos pelo modelo). De T_B para T_A, foram obtidos 44 pares de termos alinhados com precisão de 93.6% e cobertura de 95.7% (dados os 46 pares existentes e os 47 pares propostos).

Considerando dois catálogos C_A e C_B, que utilizam os tesauros T_A e T_B, respectivamente, para carregar os dados de C_A para C_B é necessário utilizar os mapeamentos de T_A para T_B, ou seja, $P(t_a, t_b)$, para que as instâncias de C_A possam ser reclassificadas utilizando os termos de T_B. Por exemplo, na Tabela 8 é possível notar que os termos classificados como “populated places” devem ser

reclassificados com “PPL”, visto que sua taxa de mapeamento é maior ou igual ao limiar 0.4.

$$P(\text{“populated places”}, \text{“PPL”}) = 0{,}9260 \geq 0{,}4$$

3.4.3 Validação e teste da abordagem adaptativa

Para definir o limiar que determina os mapeamentos corretos a partir das taxas de mapeamento calculadas pela fórmula (2), foi executado um processo de validação cruzada em seis etapas e o teste do modelo (fórmula e limiar). O procedimento de validação cruzada e teste foi aplicado tanto para os mapeamentos de T_A para T_B quanto de T_B para T_A .

Da mesma forma que a validação detalhada na seção 3.4.2, foram coletados sete conjuntos de instâncias equivalentes de C_A e C_B e separados em 6 conjuntos de afinação (*Tuning sets*) e 1 conjunto como conjunto de testes (*Testing set*). O processo de validação e teste da abordagem adaptativa segue exatamente como na seção 3.4.2.

A Figura 15 e a Figura 16 mostram os gráficos contendo os maiores índices de precisão obtidos durante a validação cruzada. Nos mapeamentos de T_A para T_B (Figura 15) obteve-se 95.4% de precisão utilizando limiar 0.4 e α = 1. Já nos mapeamentos de T_B para T_A (Figura 16) obteve-se 97.9% de precisão utilizando os mesmo valores de limiar e α. Portanto, foi definido o limiar de 0.4 e α = 1.

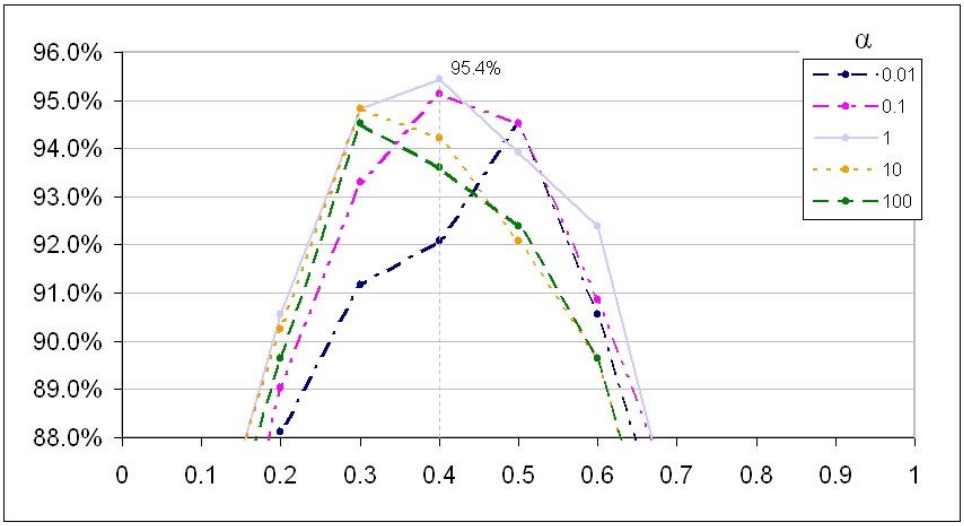

Figura 15 – Gráfico variando o α com “Precisão x Limiar” para abordagem de alinhamento *adaptativa* ($T_A \rightarrow T_B$).

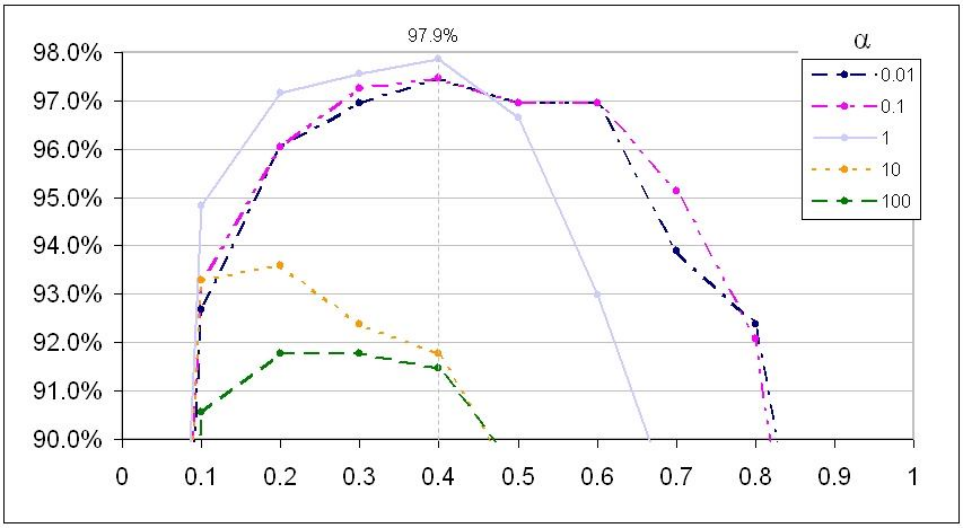

Figura 16 – Gráfico variando o α com "Precisão x Limiar" para abordagem de alinhamento *adaptativa* ($T_B \rightarrow T_A$).

De posse do limiar e de α, executou-se a etapa de teste do modelo. Nesta etapa, o conjunto de testes é primeiramente etiquetado manualmente e então submetido ao modelo. As taxas de mapeamento maiores que o limiar 0,4 são comparados às etiquetas para medir a precisão e cobertura do modelo.

Para os testes utilizamos a mesma coleção utilizada nos testes da abordagem *a priori*. A coleção contém 2.920 instâncias equivalentes entre C_A e C_B, sendo 1.998 instâncias de C_A, e 2.759 instâncias de C_B, englobando objetos que representam feições geográficas de estados da região norte do Brasil (Pará, Amapá, Roraima, Amazonas, Acre e Rondônia). Para computar as instâncias equivalentes foi utilizada uma comparação simples dos atributos de latitude e longitude, presentes em ambas as fontes utilizadas. A Tabela 7 da seção 3.4.1 mostra alguns exemplos das instâncias equivalentes coletadas.

A partir deste conjunto de dados, foram computadas as taxas de mapeamento $P(t_a, t_b)$, de T_A para T_B, e $P(t_b, t_a)$, de T_B para T_A. A Tabela 9 mostra alguns exemplos das taxas de mapeamento obtidas.

Como resultado, foram obtidos 26 pares de termos corretamente alinhados de T_A para T_B com índice de precisão de 89,7% e cobertura de 81,3% (dados os 32 pares existentes e os 29 pares propostos). De T_B para T_A, foram obtidos 44 pares de termos alinhados com precisão de 93,6% e cobertura de 95,7% (dados os 46 pares existentes e os 47 pares propostos).

Tabela 9 – Exemplos dos alinhamentos obtidos via abordagem *adaptativa*.

T_B term	T_A term	$n(t_b,t_a)$	$P(t_a,t_b)$	$P(t_b,t_a)$
administrative areas	ADM1	4	0.8003	0.8010
airport features	AIRF	7	0.5835	0.7005
educational facilities	SCH	3	0.8003	0.8010
islands	PPL	3	0.1501	0.0035
islands	ISL	213	0.9271	0.9787
lakes	LK	169	0.8795	0.9801
lakes	LGN	4	0.0268	0.8578
populated places	PPL	938	0.9373	0.9563
populated places	RSV	1	0.0012	0.0837
ridges	RDGE	12	0.4801	0.8575
ridges	SPUR	10	0.4001	0.8337
ridges	PPL	1	0.0401	0.0012
roadways	RD	2	0.6672	0.6683
tribal areas	RESV	2	0.6672	0.6683
waterfalls	FLLS	234	0.9356	0.9649

Considerando os dois catálogos C_A e C_B, que utilizam os tesauros T_A e T_B, respectivamente, dada uma consulta Q utilizando termos de T_A, para consultar dados de C_A e C_B é necessário utilizar os mapeamentos de T_A para T_B, ou seja, $P(t_a, t_b)$, para criar uma consulta Q_B para C_B, pois a consulta Q_A para C_A é facilmente criada, pois o termo utilizado na consulta Q foi de T_A.

Por exemplo, de posse das taxas de mapeamento, exemplificadas na Tabela 9, a consulta Q_A é criada utilizando o termo “populated places”; já para criar a consulta Q_B deve ser utilizado o termo “PPL”, visto que sua taxa de mapeamento é maior ou igual ao limiar 0.4.

Q = “populated places” USING T_A
*Q_A = SELECT * FROM C_A WHERE class LIKE “populated places”*
P(“populated places”, “PPL”) = 0,9373 ≥ 0,4
*Q_B = SELECT * FROM C_B WHERE class LIKE “PPL”*

Se a consulta Q_A utiliza o termo “ridges”, para criar a consulta Q_B devem ser utilizados os termos “RDGE” e “SPUR”, visto que suas taxas de mapeamento são maiores ou iguais ao limiar 0.4.

Q = “ridges” USING T_A
*Q_A = SELECT * FROM C_A WHERE class LIKE “ridges”*

P(“ridges”, “RDGE”) = 0,4801 ≥ 0,4
P(“ridges”, “SPUR”) = 0,4001 ≥ 0,4
*Q_B = SELECT * FROM C_B WHERE class LIKE “RDGE” OR class LIKE “SPUR”*

Da mesma forma, dada uma consulta *Q*, utilizando termos de T_B, para consultar dados de C_A e C_B, é necessário utilizar os mapeamentos de T_B para T_A, ou seja, $P(t_b, t_a)$, para criar uma consulta Q_A para C_A. Dada uma consulta Q_B utilizando o termo “FLLS”, para criar a consulta Q_A deve ser utilizado o termo “waterfalls”, visto que sua taxa de mapeamento é maior ou igual ao limiar 0.4.

Q = “FLLS” USING T_B
*Q_B = SELECT * FROM C_B WHERE class LIKE “FLLS”*
P(“FLLS”, “waterfalls”) = 0,9649 ≥ 0,4
*Q_A = SELECT * FROM C_A WHERE class LIKE “waterfalls”*

3.5 Considerações

Este capítulo apresentadou as abordagens para alinhamento de tesauros utilizando instâncias equivalentes como evidências.

A seção 3.2 introduziu a abordagem *a priori* para alinhamento de tesauros, onde instâncias de ambos os catálogos são previamente coletadas e analisadas, gerando evidências para os alinhamentos dos conceitos dos tesauros. A seção 3.3 introduziu a abordagem adaptativa que usa instâncias retornadas nas respostas de consultas de usuários para descoberta dos alinhamentos de forma incremental. A seção 3.4 apresentou exemplos de casos de heterogeneidade entre dois tesauros reais: do GEOnet e da ADL Gazetteer. Os tesauros foram utilizados nos experimentos apresentados nas seções 3.4.2 e 3.4.3 para validação e teste das abordagens *a priori* e adaptativa, respectivamente. Ambas as abordagens tiveram resultados satisfatórios com precisão variando entre 89.7% e 93.6%, e cobertura entre 81.3% e 95.7%. Com isso, o uso de instâncias mostrou ser uma alternativa eficaz às técnicas de alinhamento puramente sintáticas, que ignoram as divergências semânticas que podem existir entre conceitos sintaticamente similares.