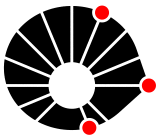

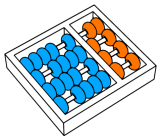

Identificação de mapeamentos entre ontologias em diferentes línguas

Gabriel Oliveira dos Santos *Julio Cesar dos Reis*

Technical Report - IC-18-09 - Relatório Técnico
July - 2018 - Julho

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE COMPUTAÇÃO

Identificação de mapeamentos entre ontologias em diferentes línguas

Gabriel Oliveira dos Santos* Julio Cesar dos Reis†

Julho 2018

Resumo

Identificação automática de mapeamentos entre ontologias em diferentes línguas envolve o estudo de medidas de similaridade. Essas medidas calculam a proximidade sintática e ou semântica entre conceitos das ontologias. Neste trabalho[1] conduzimos uma série de experimentos para avaliar abordagens de medidas de similaridade e desenvolvemos uma técnica para computar a similaridade entre dois conceitos, baseada na ponderação entre similaridades sintática e semântica. Empregamos os vetores NASARI em conjunto com a rede semântica de domínio neutro *BabelNet* para o cálculo da similaridade semântica. Resultados indicam que a atribuição de pesos nas medidas influenciam positivamente a qualidade dos mapeamentos obtidos. O efeito de combinar as medidas resulta em melhores resultados que o emprego de cada uma das formas de similaridade separadamente.

1 Introdução

Ontologias são estruturas que descrevem um domínio específico do conhecimento, de modo a relacionar os conceitos que compõem esse domínio. Essas estruturas permitem expressar a semântica de dados devido a isso têm sido largamente utilizadas na interconexão das informações entre sistemas computacionais. Mapeamentos interligam explicitamente conceitos advindos de diferentes ontologias. Eles desempenham um papel fundamental para integrar dados e outras tarefas de análise semântica.

A criação automática de mapeamentos tem sido amplamente estudada no contexto de ontologias descritas na mesma linguagem natural [11]. Entretanto, métodos para

*Instituto de Computação, Universidade Estadual de Campinas, 13081-970 Campinas, SP.
†Instituto de Computação, Universidade Estadual de Campinas, 13081-970 Campinas, SP.
[1]Pesquisa desenvolvida com suporte financeiro da FAPESP, processo #2017/23522-3

detectar relacionamento entre conceitos em diferentes línguas (no inglês, *cross-lingual matching*) ainda demandam muitos estudos. Devido ao grande volume das ontologias atuais, a investigação de métodos automáticos e precisos são essenciais para assegurar a qualidade dos mapeamentos gerados. Embora a literatura tenha apresentado técnicas de *cross-lingual matching* (*e.g.,* [2]), os estudos ainda pouco averiguaram a influência de métodos de cálculo de similaridade e características do domínio. Nesse contexto, as medidas de similaridade são chave para se obter os mapeamentos, pois permitem calcular o nível de semelhança léxica e semântica entre os conceitos.

Esta pesquisa investiga o mapeamento de ontologias descritas em diferentes linguagens naturais. Para se avaliar conceitos de ontologias em línguas distintas utilizamos duas abordagens principais: a similaridade sintática e a similaridade semântica. Entende-se similaridade sintática como sendo um valor que é calculado a partir da análise da cadeia de caracteres (extraída a partir dos rótulos de conceitos da ontologia), enquanto a similaridade semântica é calculada levando em conta informações externas à cadeia de caracteres que caracteriza os conceitos, como sinônimos e o contexto em que o termo está inserido. Ambas as medidas indicam quão similar são dois conceitos advindos de ontologias distintas (*cf.* seção 3).

Neste trabalho, desenvolvemos uma série de experimentos com o objetivo de investigar de forma empírica o impacto do uso de diferentes tipos de similaridades no mapeamento resultante entre ontologias descritas em idiomas distintos. Em nossa metodologia, comparamos um conceito de uma ontologia (em sua versão traduzida) com conceitos de outra ontologia, de forma a determinar a similaridade entre eles. Nos experimentos, traduzimos os termos para uma língua pivô (o Inglês) visando nos beneficiarmos da maior disponibilidade de recursos externos tais como dicionários, thesauri, corpora, *etc.*

Avaliamos a aplicação de diferentes técnicas de similaridade sintática e semântica em uma proposta de similaridade composta, que ponderada os valores obtidos pelas diferentes técnicas. A análise semântica se fundamenta em técnicas de domínio neutro como *Babelnet*. Nos experimentos, exploramos ontologias do domínio de conferência do *dataset MultiFarm*[2] descritas em Português, Inglês e Espanhol, criando diversos mapeamentos de conceitos descritos na língua Portuguesa, para conceitos descritos em Inglês ou Espanhol. O *MultiFarm* é um *dataset* extensamente usado para avaliar métodos de *cross-lingual matching* e apresenta mapeamentos de referência para análise de resultados. A escolha dos idiomas foi feita de modo a investigar se existe divergência considerável pelo fato de algumas palavras terem a escrita mais semelhante em Português e Espanhol, de que em Português e Inglês.

Os resultados obtidos indicam que a escolha dos pesos na similaridade composta desempenham um papel importante na acurácia dos mapeamentos. Pesos maiores para a similaridade sintática resultam em alinhamentos mais precisos. Uma possível

[2]disponível em: https://www.irit.fr/recherches/MELODI/multifarm

explicação é que o domínio de conferência possui muitos termos específicos de um domínio particular, enquanto que o tesauro *Babelnet* (usado na computação da similaridade semântica) é de um domínio neutro. Consequentemente, o alinhamento resultante das ontologias com os mapeamentos gerados se tornam menos precisos.

O restante deste relatório está organizado da seguinte forma: na seção 2 apresentamos o estado da arte; na seção 3 formalizamos os conceitos fundamentais para a proposta; a seção 4 descrevemos os experimentos realizados, bem como seus resultados e uma análise comparativa com os trabalhos relacionados; a seção 5 descreve as conclusões e os trabalhos futuros.

2 Revisão da Literatura

Propostas existentes na literatura para o problema de alinhamento de ontologias descritas em diferentes línguas visam reutilizar técnicas de alinhamento de ontologias em uma mesma língua. Essas técnicas são aplicadas no contexto de idiomas distintos, utilizando um terceiro idioma como pivô. Meilicke *et al.* [4] estudou essa abordagem, tentando aplicar técnicas de *monolingual matching* (para criar alinhamentos entre ontologias no mesmo idioma) em um *dataset* predefinido. Seus resultados indicaram dificuldades de se criar mapeamentos *cross-lingual* a partir de algoritmos de *monolingual matching* tradicionais.

Afim de lidar com o problema do *cross-lingual matching*, Trojahn *et al.* [21] desenvolveram uma série de pesquisas acerca de técnicas que endereçam o alinhamento com ontologias em línguas distintas. Eles apontaram diferentes abordagens, mas ainda com resultados preliminares. Em particular, o efeito da tradução automática tem sido estudado e os resultados indicaram que o uso de uma terceira língua no alinhamento pode ser relevante para se alcançar mapeamentos de qualidade. Embora os efeitos dessa técnica ainda precisam ser melhor estudados [8]. Nessa abordagem, para superar a barreira linguística, se faz uso de tradutores automáticos para se traduzir determinadas componentes da ontologia para uma língua pivô, normalmente o inglês, por apresentar maior recursos para consulta (dicionários, thesauri, *etc*).

Alguns trabalhos têm empregado o cálculo de similaridade de modo a melhorar os resultados do *matching* [21]. Contudo, avanços obtidos demostraram que essa abordagem ainda é pouco efetiva. Essas evidências indicam que há espaço para pesquisas que objetivam aprimorar os resultados de *matching* entre ontologias de diferentes línguas.

O uso de recursos de domínio específico ainda tem sido pouco estudado. Zhang e Bodenreider [23] propuseram o uso da estrutura de domínio biomédico *Unified Medical Language System* (UMLS)[3] como recurso externo para se criar mapeamentos entre ontologias no domínio de anatomia humana. Os resultados obtidos apontaram que o

[3]www.nlm.nih.gov/research/umls (Acesso em Maio de 2018).

uso de domínio específico é a chave para o *cross-lingual matching*, pois permite a identificação de alinhamentos adicionais em comparação ao uso de recursos independentes de domínio (genérico). Similarmente, Garla e Brandt [9] estudaram a influência de diferentes *background knowledge* do domínio biomédico na acurácia do cálculo da similaridade. Seus resultados demonstram que o uso do UMLS aliado à similaridade semântica influenciam na melhora da precisão do alinhamento de ontologias.

Em nossa revisão do estado da arte, estudamos técnicas que visam gerar automaticamente mapeamentos entre ontologias em diferentes línguas, para isso analisamos três abordagens recentes: *CroLOM* [13], *SOCOM++*[8] e o *YAM++*[7].

A proposta do *CroLOM* [13] se baseia em técnicas de processamento de linguagem natural (como *lemmatization*, *stemming* e *stopword elimination*) para normalizar os rótulos extraídos das ontologias. Em seguida, essas entidades são traduzidas para o inglês, como língua pivô. A partir disso, a técnica desenvolve um produto cartesiano com os conceitos que compõem as ontologias, que emprega de maneira híbrida as similaridades sintáticas e semântica para poder identificar alinhamentos em potencial. A similaridade sintática é calculada a partir da distância de *Levenshtein*[14], enquanto que a similaridade semântica é baseada nas categorias de palavras. Após isso, aplica-se um filtro inicial para selecionar os mapeamentos com a maior similaridade. Por fim, é aplicado um segundo filtro, que seleciona apenas os mapeamentos que possuírem uma similaridade maior que um dado limiar (*threshold*). O conjunto resultante que contém esses mapeamentos é selecionado.

A abordagem do *SOCOM++* [8] considera diversas configurações com parâmetros distintos. Diferente do *CroLOM*, nessa abordagem os rótulos dos conceitos da ontologia de origem são traduzidos para o mesmo idioma da ontologia de destino. Após isso, ambas ontologias estão descritas na mesma língua natural, então são aplicadas técnicas de alinhamento de ontologias para uma mesma língua (no inglês *monolingual-matching*). Nesse processo é avaliado o contexto de um dado conceito, que considera todos os conceitos imediatamente vizinhos a desse conceito considerado, de modo a melhorar a qualidade dos mapeamentos. Essa abordagem foi projetada para suportar ajustes nas traduções de *labels* selecionados (rótulos dos conceitos das ontologias), permitindo o usuário analisar o mapeamento resultante e decidir alterações necessárias.

Na abordagem do YAM++ [7], os rótulos dos conceitos das duas ontologias consideradas são traduzidos para a língua inglesa, como língua pivô. Em seguida, os conceitos passam por uma etapa de filtragem denominada de *candidate filtering*. Nessa etapa filtros heurísticos são aplicados para determinar mapeamentos candidatos, reduzindo o espaço de busca. Na fase seguinte, verifica-se a vizinhança dos conceitos considerados anteriormente, afim de selecionar mapeamentos mais precisos. Por fim, os mapeamentos selecionados passam por um processo de análise semântica, que remove aqueles considerados inconsistentes de modo a obter uma maior precisão nos mapeamentos resultantes.

Nossa abordagem se diferencia das propostas supracitadas, por nos basearmos na

rede semântica de domínio neutro *BabelNet* em conjunto com os vetores NASARI [18]. A similaridade semântica é calculada a partir dos métodos de *Weighted Overlap* (cf. seção 3.2.1). Adicionalmente, consideramos uma abordagem original para o cálculo da similaridade, que combina resultados de similaridades semântica e sintática, através da ponderação entre elas.

3 Fundamentos e Formalizações

Esta seção formaliza os conceitos fundamentais desta investigação.

3.1 Ontologias

Ontologias definem um vocabulário comum em um domínio, normalmente descrito como um grafo. Elas são usadas como uma representação semântica em sistemas computacionais descrevendo de forma estruturada conceitos e as relações entre eles.

Definição 3.1 (Ontologia) *Uma ontologia $\mathcal{O}$ descreve um domínio em termos de conceitos, atributos e relações [10]. Formalmente, uma ontologia $\mathcal{O} = (\mathcal{C}_\mathcal{O}, \mathcal{R}, \mathcal{A}_\mathcal{O})$ consiste em um conjunto de conceitos $\mathcal{C}_\mathcal{O}$ inter-relacionados por meio de um conjunto de relações direcionadas $\mathcal{R}$. Cada conceito $c \in \mathcal{C}_\mathcal{O}$ possui um identificador único que está associado com um conjunto de atributos $\mathcal{A}_\mathcal{O}(c) = \{a_1, a_2, ..., a_p\}$. Cada relação $r(c_1, c_2) \in \mathcal{R}$ pode ser descrita como uma tupla $(c_1, c_2, r(c_1, c_2))$, onde $r(c_1, c_2)$ é uma função que retorna o tipo da relação entre os conceitos (e.g., "$\equiv$", "$\sqsubseteq$", etc). Os símbolos "$\equiv$" e "$\sqsubseteq$" representam relações do tipo "equivalente" e "é-um", respectivamente. Adicionalmente, a relação não necessariamente precisa ser hierárquica, podendo ser uma relação do domínio. Por exemplo, no domínio biomédico, os conceitos c_1 : "Insulina" e c_2 : "Diabetes" podem estar relacionados através da seguinte função: $r(c_1, c_2) =$ "trata".*

3.2 Similaridade

Há várias maneiras de se avaliar a proximidade entre duas entidades dadas, de modo que é possível modelar tal métrica no problema desta pesquisa. Este trabalho explora os conceitos de similaridade e distância. Essas medidas, entretanto, possuem propriedades matemáticas que as definem. De acordo com Euzenat e Shvaiko [12] a similaridade é definida como:

Definição 3.2 (Similaridade) *Seja $\mathbb{S}$ um conjunto de Strings, então a similaridade é a função $sim : \mathbb{S} \times \mathbb{S} \to \mathbb{R}$ que associa cada par ordenado de strings a um número real, que expressa quão similar são as entidades, tal que:*

$$\forall x, y \in \mathbb{S}, \quad sim(x, y) \geq 0 \quad \textbf{\textit{(positividade)}}$$

$$\forall x, y, z \in \mathbb{S}, \quad sim(x, x) \geq sim(y, z) \quad \textbf{\textit{(existência de máximo)}}$$

$$\forall x, y \in \mathbb{S}, \quad sim(x, y) = sim(y, x) \quad \textbf{\textit{(simetria)}}$$

Definição 3.3 (Similaridade entre conceitos de ontologias) *Dados dois conceitos particulares c_i e c_j de uma ontologia (ou advindos de ontologias distintas), a similaridade entre eles é definida como a similaridade máxima entre os atributos de c_i e c_j. Formalmente:*

$$sim(c_i, c_j) = \arg\max \ sim(a_{ix}, a_{jy}) \tag{1}$$

considere $sim(a_{ix}, a_{jy})$ sendo a similaridade entre os pares de atributos a_{ix} e a_{jy} de c_i e c_j, respectivamente. A similaridade pode ser calculada em diferentes níveis linguísticos desde o nível de cadeia de caracteres até mesmo no nível semântico, que se refere ao significado dos termos [6].

Distância de *Levenshtein* [14], igualmente conhecida como distância de edição, é um algoritmo de similaridade sintática, que pode ser entendido como o número mínimo de edições (inserção, remoção ou troca de caracteres) sobre uma cadeira de caractere s até ela ser transformada em s'.

Definição 3.4 (Distância de *Levenshtein*) *Seja $\mathbb{S}$ o conjunto das Strings (cadeia de caracteres de entrada), $\mathbb{P}$ o conjunto dos caracteres; $\varphi : \mathbb{P} \times \mathbb{P} \rightarrow \{0, 1\}$ define a função custo de substituição de carácter e $sin : \mathbb{S} \times \mathbb{S} \rightarrow \mathbb{N}$ define a função recursiva da distância de Levenshtein, tal que:*

$$\varphi(a, b) = \begin{cases} 0, \text{se } a = b \\ 1, \text{se } a \neq b \end{cases} \tag{2}$$

Considere as cadeias de caracteres de entrada s e t como tendo tamanho, m e n, respectivamente.

$$\text{sin(s,t)} = \begin{cases} n, \text{se m} = 0 \\ m, \text{se n} = 0 \\ \min \begin{cases} sin(s[1..m], t[1..n]) + \varphi(s[0], t[0]) & \text{(troca de s[0] por t[0])} \\ sin(s[0..m], t[1..n]) + 1 & \text{(adição em s)} \\ sin(s[1..m], t[0..n]) + 1 & \text{(remoção de s[0])} \end{cases} \end{cases}$$

(3)

sin(s,t) pode ser normalizada para o intervalo $[0, 1]$, dividindo o resultado encontrado por $n + m$ [14].

3.2.1 Medidas de Similaridade Semântica

Similaridade semântica entre dois termos se refere ao quão similar são os seus respectivos sentidos (significados) em um dado contexto. Por exemplo, as palavras "manga" e "camisa" possuem uma relação muito mais próxima no contexto de vestuário, do que "manga" e "abacaxi". Por outro lado, se considerarmos o contexto de frutas, o último par de termos deve ser considerado intuitivamente mais similar que o primeiro.

Existem vários algoritmos que se propõem a calcular a similaridade semântica. Usualmente os algoritmos exploram um recurso externo como vocabulários e dicionários que auxilia no cálculo de similaridade. Neste trabalho usamos os vetores NASARI em conjunto com a rede semântica de domínio neutro *BabelNet* [18]. Para o domínio específico biomédico, exploramos o *UMLS::Similarity project* para explorar a rede semântica da *UMLS*(*Unified Medical Language System*) [16], uma rede semântica de domínio-específico biomédico que categoriza os conceitos representados no metathesaurus da UMLS.

Essa escolha nos permite desenvolver experimentos para entender a influência das medidas de similaridade semântica tanto em domínio específicos como em contextos independentes de domínio específico. NASARI apoia o cálculo de similaridade no contexto de múltiplas línguas, pois usa vetores baseados em "*synsets*" (conjuntos de sinônimos) utilizados pela *Babelnet* [1]. A partir da representação por vetores semânticos do NASARI, usamos o método de *Weighted Overlap* (cf. Equação (5)):

$$sem(el_1, el_2) = WO(v_1, v_2) \tag{4}$$

Com base na Equação (4), a similaridade entre dois elementos el_1 e el_2 (*e.g.*, termos como rótulo de conceitos de ontologias expressos em cadeia de caracteres) é dada pela função de *Weighted Overlap* com os parâmetros v_1 e v_2, que por sua vez expressam na forma de cadeia de caracteres os vetores semânticos dos elementos el_1 e el_2, respectivamente. A função de *Weighted Overlap* é definida da seguinte forma:

$$WO(v_1, v_2) = \frac{\sum_{i=1}^{|S|}(r_i^1 + r_i^2)^{-1}}{\sum_{i=1}^{|S|}(2i)^{-1}} \tag{5}$$

A função *Weighted Overlap* calcula a similaridade entre os sentidos de dois itens léxicos. Formalmente, dados dois itens léxicos v_1 e v_2, e seus respectivos *rankings* de sentidos T_1 e T_2, ordenados por grau de significância dos sentidos, definimos o

conjunto $S = T_1 \cap T_2$, de tal forma que contenha os sentidos comuns aos dois itens; r_i^j é a posição do sentido $s_i \in S$ no *ranking* T_j, onde $i = 1$ é a primeira posição (a mais alta). O somatório $\sum_{i=1}^{|S|}(r_i^1 + r_i^2)^{-1}$ define a similaridade entre os sentidos, resultando no valor máximo quando os sentidos estiverem na mesmas posições dos seus respectivos *rankings*. O denominador $\sum_{i=1}^{|S|}(2i)^{-1}$ é usado para normalizar a medida no intervalo $[0, 1]$ (Pilehvar *et. al.* [17]).

3.3 Processo de Alinhamento e Mapeamentos

Consideramos os termos mapeamento e alinhamento intercambiáveis. Entretanto os diferenciamos do conceito de *matching* (processo de alinhamento que gera os mapeamentos). Então, *matching* se refere ao processo de identificar os mapeamentos entre os conceitos, enquanto que alinhamento se refere ao resultado desse processo.

Definição 3.5 (*Alinhamento de ontologias*) *O alinhamento representa a correspondência entre dois conceitos de ontologias distintas. Formalmente, seja os conceitos $c_i \in \mathcal{C}_{\mathcal{O}_1}$ e $c_j \in \mathcal{C}_{\mathcal{O}_2}$, então o alinhamento é expresso pela tupla $m_{ci \to cj} = (c_i, c_j, r(c_i, c_j))$, onde $r(c_i, c_j) \in \mathcal{R}$ é a relação entre esses conceitos. Por exemplo, considere os conceitos $c_1 \in \mathcal{C}_{\mathcal{O}1}$ e $c_2 \in \mathcal{C}_{\mathcal{O}2}$, tal que c_1 = "Pâncreas" e c_2 = "Diabetes", o alinhamento entre esses conceitos é $m_{c1 \to c2} = (c_1, c_2, \geq)$, no qual $\geq$ indica que "Pâncreas" é um conceito semanticamente mais abrangente que "Diabetes".*

Esta pesquisa endereça o problema de alinhamento entre ontologias, cujo conteúdo está descritos em diferentes línguas naturais. Esse problema é formalizado da seguinte maneira:

Definição 3.6 (Processo de alinhamento de ontologias em diferentes línguas) *Sejam O_X e O_Y ontologias descritas nas línguas "X" e "Y", respectivamente, e $c_i \in \mathcal{C}_{\mathcal{O}_X}$ e $c_j \in \mathcal{C}_{\mathcal{O}_Y}$ os respectivos conjuntos de conceitos. O problema é identificar automaticamente o conjunto adequado de tuplas $m_{ci \to cj} = (c_i, c_j, r(c_i, c_j))$, onde $r(c_i, c_j) \in \mathcal{R}$ é a relação entre esses conceitos. Por exemplo, considere os conceitos $c_1 \in \mathcal{C}_{\mathcal{O}pt}$ e $c_2 \in \mathcal{C}_{\mathcal{O}in}$, em uma ontologia descrita em Português e em Inglês, respectivamente, tal que c_1 = "Cabeça" e c_2 = "Body", o alinhamento entre esses conceitos é $m_{c1 \to c2} = (c_1, c_2, \sqsubseteq)$. Mais precisamente, nessa investigação apenas tratamos $r(c_i, c_j) \longrightarrow \equiv$.*

Definição 3.7 (Conjunto de mapeamentos) *O resultante final do processo de alinhamento de ontologias é o conjunto que contém os mapeamentos mais adequados entre os conceitos das duas ontologias de entrada. Formalmente, o mapeamento entre as ontologias O_1 e O_2 é dado por $\mathcal{M}_{\mathcal{O}_1 \to \mathcal{O}_2}(\lambda) = \{m_{ci \to cj} | c_i \in \mathcal{C}_{\mathcal{O}_1} \wedge c_j \in \mathcal{C}_{\mathcal{O}_2} \wedge sim(c_i, c_j) \geq \lambda\}$, onde "$\lambda$"é o limiar (threshold – valor mínimo para ser considerado similar) e $sim(c_i, c_j)$ é a similaridade entre c_i e c_j.*

4 Experimentos

Este estudo visa analisar dois aspectos chaves para o problema tratado nessa investigação: (i) a efetividade das abordagens no cálculo de similaridade; e, (ii) o impacto dos idiomas das ontologias na precisão dos alinhamentos obtidos. Para esse fim, propomos uma série de três experimentos que consideram duas situações na criação dos mapeamentos: um entre os conceitos de uma ontologia descrita em Português para outra em Inglês; e outro entre os conceitos de uma ontologia descrita em Português para outra descrita na língua espanhola. Os experimentos foram conduzidos utilizando o *dataset* de referência da *MultiFarm* (*cf.* Seção 4.1), que descreve o domínio de conferência.

- **Experimento 1**: Criar os mapeamentos utilizando apenas similaridade sintática usando a técnica definida pela Equação (3) (*cf.* Algoritmo 1).
- **Experimento 2**: Criar os mapeamentos baseando-se na similaridade semântica com base na técnica definida pela Equação (4) (*cf.* Algoritmo 2).
- **Experimento 3**: Criar os mapeamentos a partir de uma nova abordagem, que leva em conta a combinação da similaridade semântica com a sintática através de uma média ponderada entre elas para determinar a similaridade entre os conceitos (*cf.* Equação (6) na seção 4.2). Esse procedimento é descrito pelo Algoritmo 3.

4.1 Materiais

Os experimentos realizados têm como base ontologias no domínio de conferência do *dataset* da *MultiFarm*[4] na versão do ano de 2015, usando majoritariamente as ontologias descritas nos idiomas Inglês e Espanhol mapeadas para Português. O *dataset* é o mesmo utilizado atualmente na *OAEI* (Ontology Alignment Evaluation Initiative)[5]

O *MultiFarm* [3] é um *benchmark* multilinguístico criado para se testar algoritmos de *matching* de ontologias em múltiplas línguas. Ele é composto por um conjunto de 7 ontologias do domínio de Conferência (Cmt, Conference, ConfOf, Edas, Ekaw, Iasted, Sigkdd) originalmente em Inglês (en), traduzidas para 8 idiomas: Chinês (cn), Tcheco (cz), Holandês (nl), Francês (fr), Alemão (de), Português (pt), Russo (ru), Espanhol (es). Os mapeamentos *cross-lingual* entre essas ontologias foram curados manualmente e podem ser considerados uma referência para a avaliação de algoritmos de alinhamentos automáticos entre ontologias.

Há alinhamentos baseados em um conjunto de 45 pares de línguas. Por exemplo, o par pt-es refere-se ao caso envolvendo o Português e o Espanhol. Para cada par há

[4]Acesso em março de 2018, https://www.irit.fr/recherches/MELODI/multifarm/

[5]http://oaei.ontologymatching.org

25 mapeamentos envolvendo as ontologias Cmt, Conference, ConfOf, Iasted e Sigkdd. As ontologias Edas e Ekaw não estão disponíveis para o uso, pois são utilizadas para os testes às cegas na competição OAEI.

4.2 Procedimentos

Investigando o conceito de similaridade, desenvolvemos uma maneira de calcular a similaridade entre dois conceitos, a partir da média ponderada entre os valores da similaridade sintática e a semântica computados com base nos rótulos dos conceitos. Nossa abordagem assume que com o uso dos pesos é possível aumentar a precisão dos alinhamentos resultantes. Os resultados são discutidos na seção 4.

Definição 4.1 (Similaridade composta) *Sejam sem(t_1, t_2) e sin(t_1, t_2) as similaridades semântica – Equação* (4)*, e sintática – Equação* (3)*, normalizadas no intervalo* $[0, 1]$*, entre os termos* t_1 *e* t_2*, respectivamente. Formalmente:*

$$simC(t_1, t_2) = \frac{\alpha sem(t_1, t_2) + \beta sin(t_1, t_2)}{\alpha + \beta} \tag{6}$$

considerando α *e* β *constantes reais.*

4.2.1 Configurações

Experimento 1. Uso da similaridade sintática (Equação (3)). O Algoritmo 1 descreve esse procedimento e os resultados estão descritos na tabela 2.

Experimento 2. Uso da similaridade semântica (Equação (4)). O Algoritmo 2 descreve esse procedimento e os resultados são apresentados na Tabela 1.

Experimento 3. Uso da similaridade composta (Equação (6)). O Algoritmo 3 descreve esse procedimento e os resultados são apresentados nas Tabelas 3 e 4.

Para cada um dos experimentos, as seguintes configurações foram usadas.

- **Configuração 1:** Mapeamentos gerados a partir das ontologias nas línguas Conference[ES]-Conference[PT].
- **Configuração 2:** Mapeamentos gerados a partir das ontologias nas línguas Conference[EN]-Conference[PT].

Para os experimentos 1 e 2 variamos os limiares de valores de similaridade nos algoritmos entre os valores $\{0.66, 0.75, 0.80, 0.95\}$, que foram escolhidos baseados nas frações $\{\frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \frac{19}{20}\}$. O limiar $0,95$ foi escolhido para avaliar o comportamento do algoritmo com um limiar próximo de $1,00$.

De modo análogo, os pesos usados no experimento 3 seguem as frações $\{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}\}$, sempre atentando para que a soma do pesos atribuídos às similaridades semântica e sintática seja igual a um.

4.2.2 Algoritmos

O Algoritmo 1 calcula o produto cartesiano do conjunto de conceitos $\mathcal{C}_{\mathcal{O}_1}$ e $\mathcal{C}_{\mathcal{O}_2}$ das ontologias $\mathcal{O}_1$ e $\mathcal{O}_2$, respectivamente. A partir de cada par de conceitos traduzidos (w_1, w_2) calcula a similaridade sintática, e verifica se ela é maior ou igual ao limiar predefinido λ. Se satisfizer essa condição, então adiciona o mapeamento $(c_1, c_2, \equiv)$, que indica que os conceitos são equivalentes, ao conjunto $\mathcal{M}_{\mathcal{O}_1 \to \mathcal{O}_2}$.

Algorithm 1: Mapeamento a partir do produto cartesiano entre ontologias utilizando distância de Levenshtein

Require: $\mathcal{O}_1, \mathcal{O}_2, \lambda \in [0, 1]$

1: $\mathcal{M}_{\mathcal{O}_1 \to \mathcal{O}_2} \leftarrow \emptyset$ {Inicializa o mapeamento com vazio}
2: **for all** $c_1 \in \mathcal{C}_{\mathcal{O}_1}$ **do**
3: $w_1 \leftarrow traduz(c_1, \text{pivô})$
4: **for all** $c_2 \in \mathcal{C}_{\mathcal{O}_2}$ **do**
5: $w_2 \leftarrow traduz(c_2, \text{pivô})$
6: **if** $sin(w_1, w_2) \geq \lambda$ **then**
7: $m_{c1 \to c2} \leftarrow (c_1, c_2, \equiv)$
8: $\mathcal{M}_{\mathcal{O}_1 \to \mathcal{O}_2} \leftarrow \mathcal{M}_{\mathcal{O}_1 \to \mathcal{O}_2} \cup \{m_{c1 \to c2}\}$
9: **end if**
10: **end for**
11: **end for**
12: **return** $\mathcal{M}_{\mathcal{O}_1 \to \mathcal{O}_2}$

O Algoritmo 2 calcula o produto cartesiano do conjunto de conceitos $\mathcal{C}_{\mathcal{O}_1}$ e $\mathcal{C}_{\mathcal{O}_2}$ das ontologias $\mathcal{O}_1$ e $\mathcal{O}_2$, respectivamente. Para cada tupla $(c_1, ling_{c_1}, c_2, ling_{c_2})$, composta pelos conceitos c_1 e c_2 e suas respectivas línguas $ling_{c_1}$ e $ling_{c_2}$, chama a função $babelnet(c_1, ling_{c_1}, c_2, ling_{c_2})$. Essa função computa a similaridade semântica, que por sua vez se baseia nos *synsets* usados pelo *Babelnet* e nos vetores semânticos do NASARI (*cf.* Seção 3.2.1), para calcular o *Weighted Overlap* (Equação (5)). Tendo calculado a similaridade semântica, o algoritmo verifica se ela é maior ou igual a um limiar predefinido λ. Se satisfizer essa condição, então adiciona o par $(c_1, c_2, \equiv)$, que indica que os conceitos são equivalentes, ao mapeamento $\mathcal{M}_{\mathcal{O}_1 \to \mathcal{O}_2}$.

Algorithm 2: Mapeamento a partir do produto cartesiano entre ontologias utilizando similaridade semântica

Require: $\mathcal{O}_1, \mathcal{O}_2 e \lambda \in [0, 1]$

$\mathcal{M}_{\mathcal{O}_1 \to \mathcal{O}_2} \leftarrow \emptyset$ {Inicializa o mapeamento com vazio}
for all $c_1 \in \mathcal{C}_{\mathcal{O}_1}$ **do**
  for all $c_2 \in \mathcal{C}_{\mathcal{O}_2}$ **do**
    $sim \leftarrow babelnet(c_1, ling_{c_1}, c_2, ling_{c_2})$
    if $sim \geq \lambda$ **then**
      $m_{c1 \to c2} \leftarrow (c_1, c_2, \equiv)$
      $\mathcal{M}_{\mathcal{O}_1 \to \mathcal{O}_2} \leftarrow \mathcal{M}_{\mathcal{O}_1 \to \mathcal{O}_2} \cup \{m_{c1 \to c2}\}$
    end if
  end for
end for
return $\mathcal{M}_{\mathcal{O}_1 \to \mathcal{O}_2}$

O Algoritmo 3 calcula o produto cartesiano do conjunto de conceitos $\mathcal{C}_{\mathcal{O}_1}$ e $\mathcal{C}_{\mathcal{O}_2}$ das ontologias $\mathcal{O}_1$ e $\mathcal{O}_2$, respectivamente. A partir de cada par de conceitos (c_1, c_2) o Algoritmo 3 calcula a similaridade semântica (precedimento descrito no Algoritmo 2); a partir das traduções dos mesmos (w_1, w_2) calcula a similaridade sintática. Com base nessas informações, computa a média ponderada, atribuindo os pesos previamente definidos α e β à similaridade sintática sim_{sin} e semântica sim_{sem}, respectivamente, resultando na similaridade composta sim_{com}. Por fim, o Algoritmo 3 verifica se a similaridade composta sim_{com} é maior ou igual a um limiar predefinido λ. Se essa condição for satisfeita, então o algoritmo adiciona o mapeamento $(c_1, c_2, \equiv)$, que indica que os conceitos são equivalentes, ao conjunto $\mathcal{M}_{\mathcal{O}_1 \to \mathcal{O}_2}$.

Algorithm 3: Mapeamento a partir do produto cartesiano entre ontologias utilizando similaridade composta

Require: $\mathcal{O}_1, \mathcal{O}_2 e \lambda \in [0, 1]$

$\mathcal{M}_{\mathcal{O}_1 \to \mathcal{O}_2} \leftarrow \emptyset$ {Inicializa o mapeamento com vazio}
for all $c_1 \in \mathcal{C}_{\mathcal{O}_1}$ **do**
  $w_1 \leftarrow traduz(c_1, \text{pivô})$
  for all $c_2 \in \mathcal{C}_{\mathcal{O}_2}$ **do**
    $w_2 \leftarrow traduz(c_2, \text{pivô})$
    $sim_{sin} \leftarrow sin(w_1, w_2)$
    $sim_{sem} \leftarrow babelnet(c_1, ling_{c_1}, c_2, ling_{c_2})$
    $sim_{com} = \frac{\alpha sim_{sin} + \beta sim_{sem}}{\alpha + \beta}$
    if $sim_{com} \geq \lambda$ **then**
      $m_{c1 \to c2} \leftarrow (c_1, c_2, \equiv)$
      $\mathcal{M}_{\mathcal{O}_1 \to \mathcal{O}_2} \leftarrow \mathcal{M}_{\mathcal{O}_1 \to \mathcal{O}_2} \cup \{m_{c1 \to c2}\}$
    end if
  end for
end for
return $\mathcal{M}_{\mathcal{O}_1 \to \mathcal{O}_2}$

Os experimentos criaram mapeamentos do tipo *Conference-Conference*, utilizando o Inglês como idioma pivô. A língua inglesa foi escolhida por possuir um maior conjunto de recursos de apoio para análise semântica de conteúdo, como dicionários e *thesauri*. Os mapeamentos resultantes dos algoritmos nas diferentes configurações foram comparados com o conjunto de mapeamentos referência do *MultiFarm* (*cf.* Seção 4.1) e medidas de precisão, cobertura e Medida-F (F-1 score) [19] foram calculadas.

4.3 Resultados

A Tabela 1 apresenta os resultados obtidos a partir dos experimentos 1 e 2 em termos das medidas de Precisão, Cobertura e Medida-F. Nos baseamos na similaridade sintática e semântica, respectivamente, para determinar o quão similar são dois conceitos (Algoritmos 1 e 2). Esses resultados consideram o mapeamento de uma ontologia descrita em Português para uma ontologia descrita em Espanhol como língua da ontologia de destino. Resultados indicam uma efetividade maior do uso da similaridade sintático ao considerar o limiar de 0,95.

Limiar de similaridade (*threshold*)	Precisão	Cobertura	Medida-F
Sintática - 0,66	0,25	0,16	0,20
Semântica - 0,66	0,21	0,12	0,15
Sintática - 0,75	0,32	0,20	0,24
Semântica - 0,75	0,29	0,09	0,14
Sintática - 0,80	0,54	0,20	0,29
Semântica - 0,80	0,35	0,10	0,15
Sintática - 0,95	0,94	0,24	0,38
Semântica - 0,95	0,48	0,11	0,18

Tabela 1: Mapeamento do Português para o Espanhol a partir similaridades semântica e sintática. MultiFarm 2015 Ontologia: Conference [ES] - Conference [PT]. Língua pivô: Inglês

A Tabela 2 apresenta os resultados obtidos a partir dos experimentos 1 e 2 com uso das medidas de similaridade sintática e semântica, respectivamente, para determinar o quão similar são dois conceitos (algoritmos 1 e 2). Esses resultados consideram o mapeamento de uma ontologia descrita em Português para uma ontologia descrita em Inglês como língua da ontologia de destino. Similar aos resultados presentes na Tabela 1, a medida de similaridade sintática obteve melhores resultados, mas com um limiar de similaridade menor (0,75).

Limiar de similaridade (*threshold*)	Precisão	Cobertura	Medida-F
Sintática - 0,66	0,24	0,15	0,18
Semântica - 0,66	0,25	0,17	0,20
Sintática - 0,75	0,48	0,29	0,36
Semântica - 0,75	0,31	0,11	0,16
Sintática - 0,80	0,61	0,21	0,31
Semântica - 0,80	0,32	0,09	0,14
Sintática - 0,95	0,78	0,15	0,24
Semântica - 0,95	0,57	0,12	0,20

Tabela 2: Mapeamento do Português para o Inglês a partir similaridades semântica e sintática. MultiFarm 2015 Ontologia: Conference [EN] - Conference [PT]. Língua pivô: Inglês

A Tabela 3 descreve os resultados obtidos a partir da execução do experimento 3, com uso da similaridade composta para determinar o quão similar são dois conceitos (algoritmo 3). Os resultados foram obtidos considerando a ontologia descrita em Espanhol como língua da ontologia de destino. Similarmente, a Tabela 4 apresenta os resultados para a ontologia destino descrito em Inglês.

Limiar de Similaridade	Peso sintático	Peso semântico	Precisão	Cobertura	Medida-F
0,66	0,50	0,50	0,49	0,15	0,23
	0,33	0,67	0,40	0,10	0,16
	0,25	0,75	0,33	0,15	0,21
	0,20	0,80	0,30	0,15	0,20
	0,67	0,33	0,69	0,30	0,42
	0,75	0,25	0,68	0,33	0,44
	0,80	0,20	0,59	0,31	0,40
0,75	0,50	0,50	0,58	0,16	0,25
	0,33	0,67	0,48	0,16	0,24
	0,25	0,75	0,45	0,18	0,25
	0,20	0,80	0,40	0,17	0,24
	0,67	0,33	0,65	0,16	0,26
	0,75	0,25	0,75	0,31	0,44
	0,80	0,20	0,72	0,33	0,45
0,80	0,50	0,50	0,65	0,16	0,26
	0,33	0,67	0,58	0,16	0,25
	0,25	0,75	0,50	0,17	0,26
	0,20	0,80	0,45	0,18	0,25
	0,67	0,33	0,65	0,16	0,26
	0,75	0,25	0,65	0,16	0,26
	0,80	0,20	0,75	0,31	0,44
0,95	0,50	0,50	0,64	0,11	0,18
	0,33	0,67	0,67	0,15	0,24
	0,25	0,75	0,69	0,16	0,26
	0,20	0,80	0,65	0,16	0,26
	0,67	0,33	0,64	0,11	0,18
	0,75	0,25	0,64	0,11	0,18
	0,80	0,20	0,64	0,11	0,18

Tabela 3: Mapeamento do Português para o Espanhol a partir da similaridade composta. MultiFarm 2015 Ontologia: Conference [ES] - Conference [PT]. Língua pivô: Inglês

Os resultado indicam que a língua da ontologia de destino, aquela para a qual se deseja mapear uma dada ontologia de origem, apresenta impactos relevantes na acurácia dos mapeamentos em função do peso sintático. Nas execuções do experimento 3 nas quais a língua da ontologia de destino é o Inglês (cf. Tabela 4), pesos maiores para a similaridade sintática provoca uma queda considerável na Medida-F quando se considera limiares de similaridade a partir de 0,75. Enquanto quando se considera o Espanhol como idioma da ontologia de destino, essa queda se dá de

maneira mais suave, evidenciando uma diferença mais acentuada quando $\lambda = 0,95$.

Uma possível explicação para esse comportamento é que a língua pivô escolhida é o Inglês, assim para os experimentos que mapeiam ontologias do Português para o Inglês é necessário apenas uma tradução, reduzindo a influência dos efeitos da tradução automática e melhorando a precisão da similaridade semântica. Nesse contexto descobrimos que atribuir maior peso a similaridade semântica (consequentemente menor peso para a similaridade sintática) resulta em mapeamentos em que há um aumento na medida-F. Ou seja, mapeamentos mais adequados são gerados.

Limiar de Similaridade	Peso sintático	Peso semântico	Precisão	Cobertura	Medida-F
0,66	0,50	0,50	0,57	0,18	0,27
	0,33	0,67	0,42	0,21	0,28
	0,25	0,75	0,32	0,18	0,23
	0,20	0,80	0,28	0,17	0,21
	0,67	0,33	0,69	0,34	0,45
	0,75	0,25	0,72	0,41	0,52
	0,80	0,20	0,68	0,21	0,32
0,75	0,50	0,50	0,60	0,17	0,26
	0,33	0,67	0,52	0,23	0,32
	0,25	0,75	0,50	0,22	0,31
	0,20	0,80	0,43	0,21	0,28
	0,67	0,33	0,58	0,21	0,31
	0,75	0,25	0,70	0,15	0,25
	0,80	0,20	0,75	0,17	0,27
0,80	0,50	0,50	0,58	0,16	0,25
	0,33	0,67	0,57	0,23	0,32
	0,25	0,75	0,52	0,23	0,32
	0,20	0,80	0,50	0,22	0,31
	0,67	0,33	0,61	0,21	0,32
	0,75	0,25	0,61	0,09	0,15
	0,80	0,20	0,73	0,15	0,25
0,95	0,50	0,50	0,64	0,19	0,29
	0,33	0,67	0,61	0,21	0,32
	0,25	0,75	0,61	0,21	0,32
	0,20	0,80	0,61	0,21	0,32
	0,67	0,33	0,64	0,19	0,29
	0,75	0,25	0,64	0,07	0,13
	0,80	0,20	0,64	0,07	0,13

Tabela 4: Mapeamento do Português para o Inglês a partir da similaridade composta. MultiFarm 2015 Ontologia: Conference [EN] - Conference [PT]. Língua pivô: Inglês

O tipo de medida de similaridade adotado interfere diretamente na acurácia dos mapeamentos resultantes, como apresenta os resultados apresentados nas tabelas 3 e 4. A similaridade composta foi a que apresentou os melhores resultados em todos os cenários considerados: experimentando a criação de mapeamentos do Português para o Espanhol, do Português para o Inglês e variando o limiar de similaridade.

Adicionalmente, a escolha dos pesos desempenha um papel fundamental nos resultados. De modo que pesos 0,75 e 0,80 para a similaridade sintática (consequentemente, 0,25 e 0,20, respectivamente, para similaridade semântica) geram os resultados melhores para a Medida-F. Todavia, o ganho da efetividade pode variar de acordo com os idiomas presentes no conteúdo das ontologias.

Os resultados detectaram uma influência do limiar de similaridade. O aumento do limiar provocou um aumento da precisão, pois nesse caso tratamos como mapeamento equivalente apenas os conceitos com alto grau de similaridade. Entretanto, a Medida-F apresentou uma queda conforme se considera limiares maiores, uma vez que altos valores do limiar desconsideram conceitos equivalentes, mas que por diversos motivos não têm uma similaridade tão alta quanto esperada tendo em vista o limiar determinado. Desta forma, a cobertura acaba tendo uma queda relevante, pois muitos mapeamentos corretos são desconsiderados, o que provoca uma redução na medida F, que é a média harmônica da precisão e a cobertura. De maneira empírica, concluímos que os limiares de similaridade que geram o mapeamentos mais acurados são $\lambda = 0,66$ e $\lambda = 0,75$, pois foram os valores que mais se destacaram no experimento 3.

4.4 Análise Comparativa com a Literatura

A Tabela 5 apresenta os resultados obtidos pelos trabalhos (sistemas de alinhamento de ontologias) que mais se destacaram na OAEI nos últimos anos. Objetivamos comparar os resultados dos nossos experimentos em relação aos sistemas dessa avaliação.

Ano	Trabalho	Precisão	Cobertura	Medida-F
2015	LogMap	0,95	0,30	0,45
2015	AML	0,93	0,50	0,64
2015	XMap	0,66	0,27	0,37
2015	CLONA	0,91	0,42	0,58
2015	LYAM++	0,26	0,15	0,19

Informações adicionais sobre esses trabalhos podem ser encontradas em [13] e no website da competição[6].

Tabela 5: Resultado na competição OAEI (Multifarm Track) sobre sistemas de alinhamento de ontologias

A Tabela 5 apresenta os resultados obtidos pelos trabalhos (sistemas de alinhamento de ontologias) que mais se destacaram na OAEI no ano de 2015. Objetiva-

mos comparar os resultados dos nossos experimentos em relação aos sistemas dessa avaliação. Analisando os resultados presentes na Tabela 5 com os resultados dos experimentos conduzidos nesta investigação, concluímos que a média da medida-F de 0,28, obtida em nossos experimentos, supera os resultados de alguns sistemas avaliados, tais como o LYAM++ [20]. Todavia, considerando o limiar de 0,66 com peso sintático igual a 0,75 (consequentemente peso 0,25 para a similaridade semântica) obtemos os melhores resultados em ambas as línguas, atingindo a marca de 0,52 e 0,44 para a medida-f em inglês e espanhol, respectivamente. Com esses resultados ultrapassamos o XMAP [22] e nos aproximamos dos dois melhores resultados o AML [5] e do CLONA [15]. Resultados apontam que escolhendo parâmetros adequados podemos conseguir mapeamentos bastante satisfatórios.

5 Considerações Finais

Alinhamento de ontologias volumosas descritas em diferentes línguas naturais se apresenta um desafio de pesquisa em aberto na literatura. Neste trabalho, conduzimos uma série de experimentos afim de estudar a aplicação do cálculo de similaridade na identificação de mapeamentos entre ontologias. Utilizamos diferentes formas de computar a similaridade entre conceitos de ontologias e propomos uma abordagem que se baseia na média ponderada das medidas de similaridade sintática e semântica. Construímos algoritmos que permitiram conduzir os experimentos e analisar a influência dos pesos atribuídos aos diferentes tipos de medida de similaridade e da escolha do limiar de similaridade nos algoritmos. Planejamos aprimorar nossa proposta de alinhamento, considerando no algoritmo o uso de diferentes recursos (*Background Knowledge*) para avaliar na similaridade semântica; averiguar no processo de alinhamento o uso da vizinhança de um dado conceito, além de considerar outras formas de calcular as similaridades sintáticas e semânticas fazendo etapas adicionais no preprocessamento da cadeia de caracteres dos rótulos que denotam os conceitos. Como trabalhos futuros, conduziremos adicionalmente novos experimentos para melhor avaliar o impacto do idioma da língua da ontologia de destino, e dos pesos da similaridade sintática, porém considerando línguas distantes do Português, tais como Alemão e o Chinês, por possuir alfabetos diferentes do usado no Português.

Agradecimentos

Este trabalho tem apoio financeiro da Fundação de Amparo à Pesquisa do Estado de São Paulo (Projeto #2017/23522-3)[7].

[7]As opiniões expressas neste trabalho não refletem necessariamente às da agência de financiamento.

Referências

[1] José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64, 2016.

[2] André O. Falcão Phillip Lord Francisco M. Couto Catia Pesquita, Daniel Faria. Semantic similarity in biomedical ontologies. *PLOS computational biology*, 5(7), 2009.

[3] Fred Freitas Willem Robert van Hage Elena Montiel-Ponsoda Ryan Ribeiro de Azevedo Heiner Stuckenschmidt Ondrej Šváb-Zamazal Vojtech Svatek Andrei Tamilin Cássia Trojahn Shenghui Wang Christian Meilicke, Raúl García-Castro. Multifarm: A benchmark for multilingual ontology matching. *Web Semantics: Science, Services and Agents on the World Wide Web*, 15, 2012.

[4] Ondrej Šváb-Zamazal Christian Meilicke, Cassia Trojahn and Dominique Ritze. In *Multilingual ontology matching evaluation–a first report on using multifarm. In The Extended Semantic Web Conference(ESWC 2012): Satellite Events*, pages 132–147. Springer, 2012.

[5] Booma S. B. Daniel F. Results of aml in oaei 2015. 2015.

[6] Duy Dinh, Julio Cesar Dos Reis, Cédric Pruski, Marcos Da Silveira, and Chantal Reynaud-Delaître. Identifying relevant concept attributes to support mapping maintenance under ontology evolution. *Web semantics: Science, services and agents on the world wide web*, 29:53–66, 2014.

[7] Zohra Bellahsene DuyHoa Ngo. Overview of yam++—(not) yet another matcher for ontology alignment task. *Web Semantics: Science, Services and Agents on the World Wide Web*, 41:30–49, 2016.

[8] Bo Fu, Rob Brennan, and Declan O'Sullivan. A configurable translation-based cross-lingual ontology mapping system to adjust mapping outcomes. *Web Semantics: Science, Services and Agents on the World Wide Web*, 15:15–36, 2012.

[9] V. N. Garla and C. Brandt. Semantic similarity in the biomedical domain: an evaluation across knowledge sources. *BMC Bioinformatics*, pages 1–13, 2012.

[10] Thomas R. Gruber. In *Toward principles for the design of ontologies used for knowledge sharing*, pages 43:907–928. International Journal of Human-Computer Studies, 1995.

[11] Pavel Shvaiko Jérôme Euzenat. In *Ontology matching: State of the art and future challenges*, pages 158–176. IEEE Transactions on Knowledge and Data Engineering, 2013.

[12] Pavel Shvaiko Jérôme Euzenat. In *Ontology Matching, 2nd Edition*, pages 85–90. Springer, Heidelberg, 2013.

[13] Abderrahmane Khiat. Crolom: Cross-lingual ontology matching system results for oaei 2017. 2017.

[14] Vladimir Iosifovich Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710, feb 1966. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

[15] Hazem Souid Mariem E. A. Clona results for oaei 2015. 2015.

[16] Bridget T McInnes, Ted Pedersen, and Serguei VS Pakhomov. Umls-interface and umls-similarity: Open source software for measuring paths and semantic similarity. In *AMIA Annual Symposium Proceedings*, pages 431–435, 2009.

[17] David Jurgens Mohammad Taher Pilehvar and Roberto Navigli. Align, disambiguate and walk: A unified approach for measuring semantic similarity. *Proceedings of ACL*, pages 1341–1351, 2013.

[18] Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construc- tion, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, pages 193:217–250, 2012.

[19] David M W Powers. Evaluation: From precision, recall and f-factor to roc, informedness, markedness correlation. *Journal of Machine Learning Technologies.*, 2:37–63, 2011.

[20] Abdel N. T. Light-weight cross-lingual ontology matching with lyam++. 2015.

[21] Cássia Trojahn, Bo Fu, Ondřej Zamazal, and Dominique Ritze. State-of-the-art in multilingual and cross-lingual ontology matching. In *Towards the Multilingual Semantic Web*, pages 119–135. Springer, 2014.

[22] Mohamed Tarek KHADIR Warith Eddine DJEDDI and Sadok BEN YAHIA. Xmap : Results for oaei 2015. 2015.

[23] Songmao Zhang and Olivier Bodenreider. Experience in aligning anatomical ontologies. *International journal on Semantic Web and information systems*, 3(2):1, 2007.