Avaliação do Uso de Métodos Baseados em LSA e WordNet para Correção de Questões Discursivas

Guilherme Passero[1,2], Aluizio Haendchen Filho[2], Rudimar Luís Scaranto Dazzi[1]

[1]Laboratório de Inteligência Aplicada (LIA)
Universidade do Vale do Itajaí (UNIVALI) – Itajaí, SC – Brazil

[2]Núcleo de Inteligência Artificial e Sistemas Inteligentes (NIASI)
Centro Universitário de Brusque (UNIFEBE) – Brusque, SC – Brazil

guilherme.passero0@gmail.com,aluizioh@terra.com.br,rudimar@univali.br

Abstract. *In this paper semantic analysis techniques for automatic short answer grading are described and evaluated. Methods based on corpus (LSA) and knowledge (WordNet) were applied. Computer scores compared well to the ones given by the teachers, with high correlation coefficients and accuracy.*

Resumo. *Neste trabalho são descritas e avaliadas técnicas de análise semântica para correção automática de questões discursivas. Foram utilizados métodos baseados em corpus (LSA) e em conhecimento (WordNet). Os escores calculados foram comparados com aqueles atribuídos pelos professores e observaram-se altos coeficientes de correlação e acurácia.*

1. Introdução

O uso de questões discursivas no processo de ensino e aprendizagem permite avaliar processos cognitivos mais elevados em comparação às questões objetivas, além de estimular as habilidades de comunicação e expressão do aluno [Burrows et al. 2015]. A correção de questões discursivas pelo professor não é uma tarefa trivial e, em turmas grandes, pode se tornar inviável. É comum que os alunos tenham que esperar dias ou mesmo semanas para receber os resultados.

Neste trabalho são comparadas duas abordagens de análise de similaridade textual para a avaliação automática de questões discursivas: (i) a análise semântica latente (LSA), uma abordagem baseada em corpus que tem apresentado resultados promissores na avaliação de questões discursivas [Santos e Favero 2015, Mohler e Mihalcea 2009]; e (ii) as métricas de similaridade entre conceitos do WordNet propostas em [Mohler e Mihalcea 2009, Wu e Palmer 1994, Leacock e Chodorow 1998, Hirst e St-Onge 1998, Resnik 1995, Lin 1998, Jiang e Conrath 1997], uma abordagem baseada em conhecimento que apresentou resultados similares às baseadas em corpus em [Mohler e Mihalcea 2009].

Na Seção 2 são apresentadas as técnicas de análise semântica utilizadas neste trabalho. A Seção 3 descreve o método empregado para coleta, processamento e análise do corpus da pesquisa e dos resultados. A Seção 4 apresenta trabalhos similares encontrados na literatura. Na Seção 5 são apresentados e discutidos os resultados obtidos. Por fim, são apresentadas as conclusões da pesquisa na Seção 6.

DOI: 10.5753/cbie.sbie.2016.1136

2. Análise de Similaridade Semântica

Nesta pesquisa foram utilizados dois modelos diferentes para medir a similaridade entre textos: o modelo LSA (*Latent Semantic Analysis*) usando como corpus a base de artigos do Wikipedia, e o modelo WordNet, descritos a seguir.

2.1. Modelo LSA

O modelo LSA (*Latent Semantic Analysis*) foi proposto por Landauer e Dutnais (1997) e consiste em uma técnica estatístico-matemática de abstração de conhecimento a partir de um corpus de textos, possibilitando a verificação de similaridade de palavras e sentenças através do seu uso contextual. A premissa é que palavras que tendem a ocorrer juntas dentro de um mesmo documento possuem alguma relação semântica.

O funcionamento do algoritmo é sintetizado a seguir, adaptado de [Landauer e Dutnais 1997].

1. *Construção da matriz termo/documento*: é uma matriz de representação do corpus utilizado, com as linhas correspondendo às palavras e as colunas aos documentos. Inicialmente, a cada entrada desta matriz é atribuído o valor da frequência absoluta de cada palavra em cada documento.
2. *Função de ajuste*: a frequência absoluta de cada palavra é ajustada por uma função considerando sua importância (ex. *log/entropy*).
3. *SVD (Singular Value Decomposition) da matriz*: a decomposição em valores singulares evidencia as correlações entre palavras.
4. *Redução para o espaço semântico*: a matriz é reduzida para uma dimensão entre 300 e 500 para eliminar as linhas e colunas com os menores valores singulares.

Após isso, um vetor de representação semântica de um dado conjunto de palavras pode ser obtido e comparado com outros vetores semânticos. Normalmente essa comparação é realizada calculando-se o cosseno do ângulo entre os vetores.

2.2. Modelo WordNet

O WordNet começou com um projeto de pesquisa da Princeton University [Fellbaum 1998] e pode ser visto como uma base de conhecimento onde substantivos, verbos, advérbios e adjetivos são organizados por uma variedade de relações semânticas. As palavras do léxico são mantidas dentro de um ou mais conjuntos de sinônimos (*synsets*), que representam conceitos. Como um dicionário comum, o WordNet contém as definições de palavras, mas difere porque ao invés de ser organizado alfabeticamente, é organizado conceitualmente [Leacock e Chodorow 1998].

Alguns exemplos de relações semânticas usadas pelo WordNet são hipernímia/hiponímia (é-um), meronímia (é-parte-de), sinonímia e antonímia. Essas relações são associadas com palavras para formar uma estrutura hierárquica, que é uma ferramenta útil para a linguística computacional e processamento de linguagem natural [Meng et al. 2013].

A Figura 1 apresenta o conceito “cachorro” (*synset* 02084071-n) e alguns dos seus relacionamentos no WordNet.

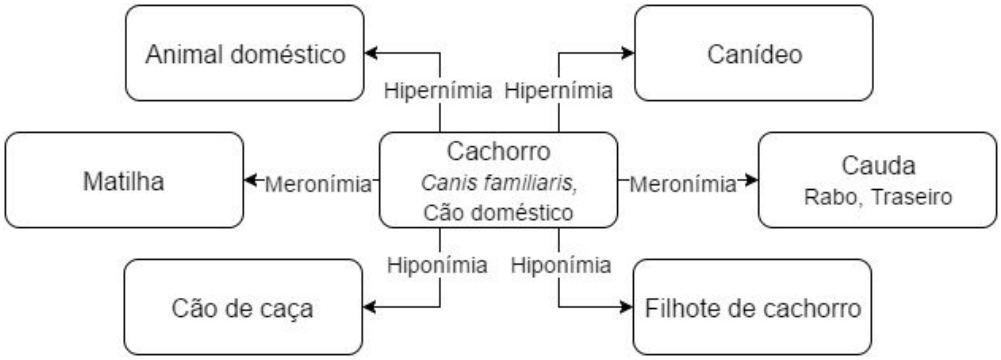

Figura 1. Conceito "cachorro" e alguns relacionamentos no WordNet

Oliveira et al. (2015) comparam sete *wordnets* disponíveis para a língua portuguesa. Segundo os autores, o OWN-PT (OpenWordNet-PT) tem conteúdo livre, é mantido com técnicas de aprendizado de máquina e revisão humana colaborativa e tem se destacado por ter sido adotado como WordNet do português pelos projetos FreeLing, Open Multilingual Wordnet e Google Translate. Considerando não haver um método de avaliação preciso para determinar o melhor WordNet para um contexto, este estudo optou pela instância mais popular, o OWN-PT, apresentado em [Paiva et al. 2012].

Várias formas de medir a similaridade têm sido propostas. As abordagens utilizadas nesta pesquisa foram as métricas *path-based* e *information content-based*.

I - Métricas *path-based*

A ideia das medidas *path-based* é que a semelhança entre dois conceitos é uma função baseada na distância entre eles e seu posicionamento no WordNet. A medida mais simples é a do caminho mais curto (*shortest path*), que considera a menor distância entre dois conceitos. A fórmula proposta por [Mohler e Mihalcea 2009] é :

$$similaridade_{shortest\_path} = \frac{1}{menor\_distancia}$$

Wu e Palmer (1994) introduziram uma escala de medida de similaridade que assume a posição dos conceitos c_1 e c_2 na taxonomia em relação à posição do conceito específico mais comum (LCS ou *lower common subsummer*). A fórmula proposta é:

$$similaridade_{W\&P} = \frac{2 * profundidade(LCS)}{profundidade(conceito_1) + profundidade(conceito_2)}$$

A métrica proposta por Hirst e St-Onge (1998) classifica as relações entre conceitos no WordNet pela direção (ex. hipônimos seriam uma relação para-baixo). Caminhos curtos e que não mudam de direção com frequência são considerados como indicadores de forte relação semântica entre conceitos [Pedersen e Michelizzi 2004]. Essa métrica aplica a fórmula:

$$similaridade_{H\&S} = C - distancia(conceito_1, conceito_2) - k * md(conceito_1, conceito_2)$$

Onde C e k são constantes e a função md retorna o número de mudanças de direção no caminho entre os conceitos. Neste trabalho foram considerados os valores praticados em [Hirst e St-Onge 1998] para essas constantes (C=8, k=1).

II - Métricas information content-based

Em medidas de similaridade baseadas em *information content* (IC) são inferidas probabilidades de conceitos ocorrerem em um corpus. O IC de um conceito pode ser quantificado como o log negativo da probabilidade de ocorrência, o que significa que à medida que a probabilidade aumenta, o IC diminui [Meng et al. 2013].

$$IC(conceito) = -log\, P(conceito)$$

A probabilidade considera tanto a ocorrência do próprio conceito quanto de seus hipônimos. Assim, quanto mais abstrato o conceito é, maior sua probabilidade associada e menor o seu *information content*.

Resnik (1995) propôs o uso do *information content* do conceito específico mais comum entre dois conceitos para indicar a sua similaridade.

$$similaridade_{Resnik} = IC(LCS)$$

Lin (1998) e Jiang e Conrath (1997) adaptaram a proposta de Resnik para considerar também o *information content* dos conceitos comparados, com as fórmulas:

$$similaridade_{Lin} = \frac{2 * IC(LCS)}{IC(conceito_1) + IC(conceito_2)}$$

$$similaridade_{J\&C} = \frac{1}{IC(conceito_1) + IC(conceito_2) - 2 * IC(LCS)}$$

3. Abordagem Proposta

A abordagem proposta compreende três etapas: (i) Coleta e Armazenamento das Informações; (ii) Pré-processamento; (iii) Processamento e Análise das Informações.

3.1. Coleta e Armazenamento dos Dados

Para a coleta, armazenamento e processamento dos dados, foi utilizado um software aplicativo próprio. As questões foram elaboradas por 14 professores do Ensino Médio de uma escola situada em Brusque/SC, com base no conteúdo visto em aula no 1º ano, e foram respondidas por alunos do 1º ao 3º ano. Os procedimentos para a coleta e armazenamento de questões e respostas foram os seguintes:

1. Apresentação de um seminário para os professores da escola com um especialista tratando sobre boas práticas na formulação de questões;
2. No mesmo encontro os professores foram orientados para a formulação de um conjunto de duas a quatro questões, dentro de suas respectivas áreas de atuação, e para a entrada dessas questões no software aplicativo desenvolvido;
3. Foram formuladas 45 questões, que foram analisadas e validadas pela especialista palestrante;
4. Foram selecionadas para compor uma prova 21 questões de nove áreas de conhecimento, visando que a realização não ultrapassasse duas horas;
5. As questões foram respondidas por 27 alunos perfazendo um total de 477 respostas, desconsiderando as que ficaram em branco.

Para o corpus desta pesquisa foram selecionadas três questões discursivas, duas da disciplina de Língua Portuguesa e uma de Geografia. Foram coletadas 27, 24 e 25 respostas para as questões 1, 2 e 3, totalizando 76 respostas, e o número médio de palavras por resposta foi 30, 11 e 30, respectivamente. A cada resposta foi atribuída uma nota de 0 a 10 por dois professores especialistas na área. As questões e respostas de referência do professor são apresentadas na Tabela 1.

Tabela 1. Questões utilizadas no corpus da pesquisa

Geografia	**Questão 1** **Enunciado**: Explique a função do horário de verão no Brasil. **Resposta esperada**: Economizar energia aproveitando melhor a luz, que se estende por mais tempo principalmente no sul do Brasil. Esta região está mais voltada para o sol devido ao eixo de inclinação da Terra.
Português **Contextualização** Leia o texto abaixo. Marcos, 31 anos, foi preso na tarde deste domingo, 31, em Brusque, com um mandado de prisão ativo, do Tribunal de Justiça do Paraná. Os policiais do Pelotão de Patrulhamento Tático (PPT) estavam em rondas quando viram Marcos pilotando uma motocicleta. Ao consultar o sistema, constataram o mandado de prisão e também que o veículo estava em situação irregular, o qual foi guinchado e conduzido ao pátio do guincho. Marcos foi conduzido ainda durante a tarde para a Unidade Prisional Avançada (UPA) de Brusque.	**Questão 2** **Enunciado**: Identifique a voz verbal que se sobressai no texto. **Resposta esperada**: Voz passiva. **Questão 3** **Enunciado**: Justifique o uso da voz verbal predominante. **Resposta esperada**: Voz passiva; destaque na ação que ocorreu; generalização do sujeito.

Em alguns casos o professor abordou na resposta esperada além do que era pedido na questão. Nesses casos a resposta foi ajustada, de modo a tratar concisamente sobre o problema apresentado. Ainda, após a validação, as respostas de referência foram resumidas mantendo sua representatividade, onde "A voz predominante é a voz passiva", por exemplo, reduziu para "Voz passiva".

3.2. Pré-processamento

Após ter sido coletado e armazenado, o corpus de respostas foi submetido a procedimentos de depuração utilizando o CoGrOO 4 [Colen 2013], com a tokenização, identificação de nomes próprios, identificação de partes do discurso (*pos-tagging*), lematização e remoção de *stopwords*. A ortografia das respostas foi revisada manualmente para garantir o correto funcionamento do CoGrOO.

A base completa de artigos do Wikipedia (versão de maio/2016) foi traduzida por um *parser* XML para um formato somente texto, mantendo a divisão dos 1,4 milhão de artigos. O mesmo procedimento de depuração aplicado às respostas foi aplicado ao Wikipedia. Nesse processo o tamanho da base passou de cerca de 5,5GB para 4GB.

3.3. Processamento e Análise das Informações

Na fase de processamento e análise os métodos de análise de similaridade semântica estudados foram executados a fim de aferir a similaridade entre a resposta do aluno e a resposta de referência do professor.

Foram criados modelos LSA com a base pré-processada do Wikipedia utilizando a biblioteca aberta Semantic Vectors [Widdows e Ferraro 2008], com as dimensões

[200, 250, 300, 350, 400, 450, 500] e os 441 mil termos mais frequentes (frequência >= 10). O modelo com dimensão 350 foi selecionado para representar o LSA neste estudo, dado que apresentou os melhores resultados.

O framework Apache Jena foi utilizado para carregar os dados do OWN-PT para a memória principal utilizando SPARQL (*SPARQL Protocol and RDF Query Language*). Os algoritmos usados para cálculo da similaridade foram adaptados a partir da biblioteca livre WordNet::Similarity [Pedersen e Michelizzi 2004].

Para o modelo WordNet o índice de similaridade foi calculado considerando um vetor com o tamanho da resposta de referência pré-processada, aplicando a fórmula:

$$similaridade_{resposta} = \frac{\sum_{i=1}^{n} max(funcao_x(palavra\_referencia_i, palavras\_resposta))}{tamanho\_resposta\_referencia}$$

Cada palavra da resposta de referência ($palavra\_referencia_i$) foi comparada com todas as palavras da resposta do aluno pré-processada (*palavras_resposta*), e a maior similaridade encontrada preencheu uma posição do vetor. Ao final, obteve-se a média aritmética dos valores desse vetor. Esse processo foi repetido para cada método estudado ($funcao_x$). Na ausência de similaridade, a distância Levenshtein foi utilizada. Como as técnicas consideravam somente substantivos e verbos − exceto a de Hirst e St-Onge −, os adjetivos da resposta esperada foram considerados somente quando encontrados na resposta do aluno.

Os índices de similaridade encontrados foram normalizados para o intervalo 0 a 10 para o cálculo de acurácia e correlação, o mesmo intervalo que fora utilizado pelos avaliadores humanos. No cálculo da acurácia utilizamos a fórmula de [Santos e Favero 2015], adaptada para o valor máximo 10:

$$acuracia = \frac{10 - erro\_medio}{10} * 100$$

4. Trabalhos Relacionados

A seguir são apresentados alguns trabalhos encontrados na literatura com abordagens similares para a correção automática de questões discursivas.

Em [Ávila e Soares 2013] foram aplicadas técnicas de pré-processamento textual e comparação de *strings* em respostas de alunos do ensino superior. Os autores otimizaram os resultados usando a distância Levenshtein e técnicas de pré-processamento textual como remoção de *stopwords*, *stemming* e troca de palavras semelhantes (dicionário mantido pelo usuário).

Este trabalho estende [Ávila e Soares 2013], com algumas diferenças no pré-processamento, principalmente em relação ao uso do CoGrOO para análise morfológica e o uso de lematização ao invés de *stemming*. Ainda, em acordo com a sugestão de extensão dos autores, foi utilizado LSA e um dicionário público (WordNet).

Em [Santos e Favero 2015] foi aplicada uma versão melhorada da técnica LSA em um corpus de 359 respostas a duas questões de vestibular de uma universidade pública. Os autores obtiveram uma acurácia de 84,94%, semelhante aos 84,93% de concordância entre os avaliadores humanos.

Mohler e Mihalcea (2009) exploraram técnicas baseadas em corpus e em conhecimento sobre 21 questões e 637 respostas na língua inglesa de estudantes de Ciências da Computação. Os melhores resultados foram obtidos utilizando LSA com um corpus de artigos do Wikipedia pertinentes ao domínio e um refinamento baseado nas melhores respostas (r=0,5099*)*. Utilizando o WordNet com o método do caminho mais curto (*shortest path*) obteve-se resultado similar (r=0.4887). No tópico a seguir é apresentada uma comparação entre os resultados alcançados por Mohler e Mihalcea e os obtidos neste trabalho.

Em consulta à literatura não encontramos trabalhos sobre análise de similaridade textual utilizando um WordNet em português.

5. Resultados e Discussão

A concordância entre os avaliadores (*judge agreement rate*) foi calculada usando a correlação de Pearson (r) e acurácia, obtendo-se os valores apresentados na Tabela 2.

Tabela 2. Concordância entre os avaliadores das respostas do corpus utilizado

Questão	Correlação (r)	Acurácia (%)
1	0,82	82,08
2	0,84	85,60
3	0,71	76,30
Total	0,70	81,18

Também foi observado que os avaliadores forneceram a mesma nota em 27 respostas (35,53%), diferenciaram em um ou dois pontos em 25 (32,89%), três a cinco em 21 (27,65%) e seis a dez pontos em 3 (3,95%).

A Tabela 3 apresenta os resultados obtidos com os métodos de similaridade utilizados em comparação às notas atribuídas pelos avaliadores. Os métodos foram agrupados por *corpus-based* (c-b), *path-based* (p-b) e *information content-based* (ic-b).

Tabela 3. Sumário dos resultados obtidos com os métodos utilizados

		Questão							
		1		2		3		Total	
Tipo	Método	Correl. (r)	Acurácia (%)	Correl. (r)	Acurácia (%)	Correl. (r)	Acurácia (%)	Correl. (r)	Acurácia (%)
c-b	LSA	0,86	85,93	0,54	67,60	0,89	88,33	0,67	80,66
p-b	*Shortest path*	0,75	75,19	0,85	79,60	0,86	90,83	0,74	81,58
	Wu e Palmer	0,75	82,96	0,81	76,80	0,84	79,17	0,77	79,74
	Leacock e Chodorow	0,66	81,11	0,81	76,40	0,82	82,50	0,75	80,00
	Hirst e St-Onge	0,71	73,33	0,82	71,20	0,79	82,50	0,74	75,53
ic-b	Lin	0,71	66,67	0,87	81,20	0,82	81,25	0,70	76,05
	Resnik	0,68	62,59	0,87	81,20	0,78	80,42	0,66	74,34
	Jiang e Conrath	0,68	62,59	0,87	81,20	0,78	80,42	0,66	74,34

Os totais apresentados na Tabela 3 indicam que as métricas utilizadas tiveram correlação semelhante à concordância entre humanos. As métricas baseadas em caminho tiveram a maior correlação (0,74-0,77) e o método do caminho mais curto a maior acurácia (81,58%).

A questão 1 tinha uma resposta de referência explicativa de 32 palavras. Para essa questão, o modelo LSA apresentou a maior acurácia e correlação (85,93% e 0.86). Por outro lado, a questão 2 tinha a menor resposta de referência, exigindo do aluno citar a expressão "voz passiva" na sua resposta. Nessa questão o LSA teve performance inferior ao modelo WordNet, que atingiu a acurácia 81,20% e correlação 0,87 com técnicas baseadas em *information content*. A técnica do caminho mais curto teve resultado similar (acurácia=79,60% e r=0,85). Para a questão 3 o método do caminho mais curto apresentou a melhor acurácia (90,83% e r=0,86), mas o modelo LSA teve a melhor correlação (r=0,89).

Na Tabela 3 uma maior correlação indica maior proximidade entre a distribuição do conjunto de notas atribuídas por humanos e o conjunto das notas computadas. Uma correlação perfeita (r=1) indicaria que todas as notas dadas pelo professor têm a mesma distribuição das notas dadas pelo computador, mas não significa que o computador atribuiu a mesma nota que o professor. Por exemplo, dado os conjuntos de notas [4, 5, 5] e [8,10,10], a correlação entre eles seria perfeita, mas a acurácia seria 53,33%. No entanto, supondo que o segundo conjunto tenha sido gerado por um algoritmo, um refinamento para dividir qualquer nota dada por dois manteria a correlação e aumentaria a acurácia para 100%. Assim, apesar da técnica do caminho mais curto ter apresentado a melhor acurácia total, a inserção de uma função de ajuste nas técnicas com maior correlação poderia resultar em uma acurácia maior.

Foram executados testes com as respostas fornecidas pelos professores antes da validação e ajuste pela especialista e foi constatada uma redução de até 12% na acurácia para o modelo WordNet. Isso pode ser explicado devido às métricas baseadas em WordNet exigirem que todos os conceitos esperados sejam abordados na resposta do aluno para um índice de 100%. Desse modo, ajustando a resposta "A voz predominante é a voz passiva" para "Voz passiva", alunos que não incluíssem uma palavra relacionada ao conceito "predominar" não seriam prejudicados. Para o modelo LSA, as respostas não ajustadas também resultaram em acurácia e correlação reduzidas (-2% e -0.13).

A Tabela 4 compara os resultados obtidos com os apresentados por Mohler e Mihalcea (2009). Para isso, os totais da Tabela 3 foram ajustados considerando um intervalo de notas entre 0 a 5, o mesmo utilizado por esses autores. Mohler e Mihalcea apresentaram vários resultados utilizando o modelo LSA, mas na Tabela 4 é apresentada a correlação referente ao uso de uma base completa de artigos do Wikipedia.

Tabela 4. Comparação dos resultados obtidos (i) com [Mohler e Mihalcea 2009] (ii)

Método	(i) (r)	(ii) (r)
LSA	0,6481	0,4286
Shortest path	0,7084	0,4413
Wu e Palmer	0,6816	0,3366
Leacock e Chodorow	0,6169	0,2231
Hirst e St-Onge	0,6604	0,1961
Lin	0,6533	0,3916
Resnik	0,6227	0,2520
Jiang e Conrath	0,6227	0,4499

É possível que a diferença dos resultados obtidos nos nossos experimentos em relação aos de Mohler e Mihalcea (2009) seja devido: (i) a diferenças nas técnicas de pré-processamento das respostas e da base do Wikipedia; (ii) a diferenças no tamanho e conteúdo da versão do Wikipedia (nós português, eles inglês); (iii) à diferença entre o grau de complexidade das questões; (iv) à forma como a resposta de referência foi elaborada. A normalização para o intervalo 0 a 5 reduziu a correlação devido à intensificação da dispersão das notas em relação à tendência linear. Assim, dado os conjuntos de notas [10, 8, 4, 3] e [9, 7, 5, 2] (r=0.95), a normalização para [5, 4, 2, 2] e [5, 4, 3, 1] (r=0,88) reduziria a correlação em 0,07. Haja vista que a maioria dos trabalhos na literatura usam escalas menores que 10 [Burrows et al. 2015], torna-se importante o estudo de outros critérios que junto ao coeficiente de correlação facilitem a comparação dos resultados entre trabalhos.

6. Conclusão

Nesta pesquisa foi avaliado o uso de métricas existentes de similaridade semântica na correção automática de questões discursivas. As principais contribuições deste trabalho são (i) a introdução ao uso de um WordNet em português na avaliação automática de questões discursivas, (ii) a constatação da variabilidade dos resultados obtidos em função do tipo da questão e do contexto de aplicação, (iii) a verificação da importância da representatividade e concisão da resposta de referência para a efetividade das técnicas avaliadas, e (iv) a crítica ao uso da correlação para comparação entre estudos, dado que essa métrica varia de acordo com o intervalo de valores utilizado e não indica necessariamente uma maior acurácia.

Referências

Ávila, R. L. F. e Soares, J. M. (2013). Uso de técnicas de pré-processamento textual e algoritmos de comparação como suporte à correção de questões dissertativas: experimentos, análises e contribuições. Anais do Simpósio Brasileiro de Informática na Educação (SBIE 2013),

Burrows, S., Gurevych, I. e Stein, B. (2015). The eras and trends of automatic short answer grading. International Journal of Artificial Intelligence in Education, v. 25, n. 1, p. 60–117.

Colen, W. (2013). Aprimorando o corretor gramatical CoGrOO. Dissertação de mestrado. Universidade de São Paulo.

Fellbaum, C. (1998). WordNet: An electronic lexical database, Language, Speech, and Communication. MIT Press.

Hirst, G. e St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. WordNet - An Electronic Lexical Database, n. April, p. 305–332.

Jiang, J. J. e Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. Rocling X.

Landauer, T. K. e Dutnais, S. T. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. Psychological Review, 104(2), 211–240.

Leacock, C. e Chodorow, M. (1998). Combining local context and WordNet sense similarity for word sense identification. WordNet: an electronic lexical database, language, speech, and communication (c. 11, p. 265–284).

Lin, D. (1998). An Information-Theoretic Definition of Similarity. Proceedings of ICML, p. 296–304.

Meng, L., Huang, R. e Gu, J. (2013). A Review of Semantic Similarity Measures in WordNet. International Journal of Hybrid Information Technology, v. 6, n. 1, p. 1–12.

Mohler, M. e Mihalcea, R. (2009). Text-to-text Semantic Similarity for Automatic Short Answer Grading. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09), April, p. 567–575.

Oliveira, H. G., Paiva, V., Freitas, C., Rademaker, A., Real, L. e Simões, A. (2015). As wordnets do português. Oslo Studies in Language, v. 7, n. 1, p. 397–424.

Pedersen, T., Patwardhan, S. e Michelizzi, J. (2004). WordNet::Similarity: Measuring the Relatedness of Concepts. In Demonstration Papers at HLT-NAACL 2004. HLT-NAACL-Demonstrations'04. Association for Computational Linguistics.

Resnik P. (1995). Using information content to evaluate semantic similarity. Proceedings of the 14th International Joint Conference on Artificial Intelligence, August 1995, p. 20-25.

Santos, J. C. A. e Favero, E. L. (2015). Practical use of a latent semantic analysis (LSA) model for automatic evaluation of written answers. Journal of the Brazilian Computer Society, v. 21, n. 1, p. 21.

Paiva, V., Rademaker, A. e Melo, G. (2012). OpenWordNet-PT: An Open Brazilian Wordnet for Reasoning. In Proceedings of COLING 2012: Demonstration Papers (pp. 353–360). Mumbai, India: The COLING 2012 Organizing Committee.

Widdows, D. e Ferraro, K. (2008). Semantic Vectors: a Scalable Open Source Package and Online Technology Management Application. LREC'08 Proceedings of the Sixth International Conference on Language Resources and Evaluation, January 2008, p. 1183–1190.

Wu Z. e Palmer M. (1994). Verb semantics and lexical selection. Proceedings of 32nd annual Meeting of the Association for Computational Linguistics, June 1994 p. 27-30.

Ziai, R., Ott, N. e Meurers, D. (2012). Short Answer Assessment: Establishing Links Between Research Strands. Proceedings of the Seventh Workshop on Building Educational Applications Using NLP. Association for Computational Linguistics.