NEPaLE: Uma ferramenta computacional de suporte à avaliação de paráfrases

Rafael de Oliveira Teixeira[1], Eloize Rossi Marques Seno[1], Helena de Medeiros Caseli[2]

[1]Instituto Federal de São Paulo – câmpus São Carlos
Rodovia Washington Luís, km 235 SP-310 prédio AT-6 Bairro Monjolinho
CEP 13.565-905 São Carlos- SP

[2]Departamento de Computação – Universidade Federal de São Carlos
Rodovia Washington Luís, km 235 SP-310 Bairro Monjolinho CEP 13.565-905

rafael.ot@outlook.com, eloize@ifsp.edu.br, helenacaseli@dc.ufscar.br

1. Introdução

Paráfrases são formas alternativas de se comunicar uma mesma mensagem (ou informação). Por exemplo, pode-se dizer que *fazer pesquisa* é "reunir e investigar dados sobre um determinado assunto" ou, ainda, "levantar e analisar dados a respeito de um assunto". Nos últimos anos, o reconhecimento automático de paráfrases, uma das subáreas do Processamento de Língua Natural (PLN), tem sido foco de muitas pesquisas (Androutsopoulos and Malakasiotis, 2010; Bannard and Callison-Burch, 2005; Barzilay and McKeown, 2001; Zhao et al., 2009). Algoritmos capazes de reconhecer paráfrases em textos são desejáveis em várias aplicações, tais como em sistemas de perguntas e respostas e de recuperação de informações na *web*, para recuperar respostas ou documentos relevantes que contêm paráfrases dos termos usados na pergunta/consulta do usuário (vide, por exemplo, Duboue and Chu-Carrol, 2006 e Riezler et al., 2007), em tradutores automáticos, para melhorar a qualidade das sentenças traduzidas automaticamente (vide, por exemplo, Barreiro, 2008), em sistemas de sumarização multidocumento, para eliminar a redundância de informações nos sumários (vide, por exemplo, Barzilay and McKeown, 2005), entre outras.

Apesar da vasta quantidade de métodos de reconhecimento e extração de paráfrases disponíveis na literatura, não se tem conhecimento de ferramentas computacionais que auxiliem no processo de avaliação dos resultados produzidos por esses métodos. Em geral, a avaliação das paráfrases extraídas automaticamente é feita por humanos, uma tarefa que consiste em analisar pares de candidatas a paráfrases, considerando um contexto bem definido, com o intuito de verificar se a relação de paráfrases se confirma ou não naquele contexto (vide, por exemplo, Bannard and Callison-Burch, 2005 e Barzilay and Mckeown, 2005). Essa tarefa pode demandar várias horas de trabalho, dependendo do tamanho do conjunto de dados a ser avaliado.

Dado esse contexto, este artigo descreve a construção de uma ferramenta de suporte à avaliação de paráfrases. Essa ferramenta, denominada NEPaLE (*Never-Ending Paraphrase Learner Evaluation*), está sendo desenvolvida no contexto do projeto "Aprendendo com a *web* a traduzir e parafrasear textos" (2013/11811-0 da

FAPESP), que tem por finalidade o aprendizado automático sem fim de conhecimento linguístico a partir de páginas da *web,* por meio do desenvolvimento de vários aprendizes automáticos, entre eles o NEPaL (*Never-Ending Paraphrase Learner*)[1]. Desse modo, um dos recursos gerados com o referido projeto é o primeiro repositório de paráfrases em nível de palavras (isto é, paráfrases lexicais como *bujão* e *botijão*) e de sintagmas (por exemplo, *capital paulista* e *capital de São Paulo*) do português do Brasil do qual se tem conhecimento. O conceito de paráfrases adotado neste trabalho se refere à duas sequências distintas de palavras que em determinados contextos podem ser intercambiáveis sem alterar o sentido original.

Mais especificamente, pretende-se com essa ferramenta contribuir para a avaliação das paráfrases obtidas pelos métodos desenvolvidos no projeto da FAPESP, minimizando o tempo despendido pelos avaliadores nessa tarefa e tornando o processo mais homogêneo e padronizado. Além disso, essa ferramenta é fundamental também para a construção do conjunto de dados de treinamento e teste do algoritmo de aprendizado sem fim de paráfrases. Ainda, dado que as paráfrases no nível lexical em muitos casos podem ser tratadas sinônimos, a NEPaLE também pode ser útil para a avaliação/anotação de sinônimos.

2. A Ferramenta NEPaLE

NEPaLE é uma ferramenta visual *online* que está sendo desenvolvida para possibilitar a avaliação *online* de paráfrases. Para o desenvolvimento da ferramenta tem sido utilizada a linguagem PHP.

O arquivo de entrada da NEPaLE (isto é, o arquivo de avaliação) é carregado na ferramenta pelo próprio avaliador e representa o conjunto de dados contendo as candidatas a paráfrases e o contexto no qual elas serão analisadas (por exemplo, um parágrafo ou sentença exemplo em que qualquer uma das candidatas ocorre). O arquivo de avaliação deve estar no formato *arff*, conforme exemplo apresentado na Figura 1. O *arff* é o formato de arquivo usado pelo *toolkit* de mineração de dados WEKA (Witten and Hall, 2005), muito utilizado na área e que implementa diversos tipos de algoritmos de Aprendizado de Máquina. A escolha desse modelo de arquivo possibilita que o conjunto de dados analisado pelo humano seja utilizado para treinamento e teste dos algoritmos disponíveis na WEKA.

Conforme se pode ver em destaque na Figura 1, o par de candidatas a paráfrases e o contexto são representados no *arff* iniciando a linha com o símbolo %[2]. No exemplo em destaque *cand =1183* indica o número do par de candidatas a paráfrases, *eloize* é o nome do avaliador[3], e *apontar* e *identificar* representam o par de possíveis paráfrases. Na linha que segue, está o contexto a ser considerado no julgamento daquele par de candidatas a paráfrases, no qual os números *118 118* têm por finalidade informar em que posição da sentença (ou texto) de exemplo ocorreu uma das candidatas do par, pois

1 Para mais informações sobre o projeto "Aprendendo com a web a traduzir e parafrasear textos" (FAPESP 2013/11811-0) acesse http://www.lalic.dc.ufscar.br/never-ending/.

2 O símbolo ‘%’ indica para os algoritmos da WEKA que aquela linha está comentada e deve ser ignorada no processamento.

3 Cada avaliador é identificado pela ferramenta através de um login individual.

durante a avaliação ela será destacada pela ferramenta de modo a facilitar a análise pelo humano (vide destaque em amarelo na Figura 2).

```
% cand=1183 eloize apontar identificar
% 118 118 seus críticos dizem que você teria uma agenda secreta no debate sobre legalização das drogas . quais são seus interesses nesta questão ?
acho_que dependência de drogas é um problema insolúvel porque , de alguma maneira , é inerente a natureza humana . nem todo_mundo se torna
dependente de drogas , mas algumas pessoas , sim . e eu não conheço a solução para isso , mas sei que a guerra as drogas , que trata aqueles que
sofrem de dependência como criminosos , tem causado mais danos do que a dependência em si . um dos objetivos da fundação é o que chamamos de
redução de danos . neste caso , então , esforços têm sido apontar os efeitos danosos da guerra as drogas .
apontar,name,identificar,esforço,ter,ser,o,efeito,danoso,v,v,n,v,v,det,n,adj,0.182,0.799,yes

% cand=1184 eloize apontar identificar
% 24 24 os investidores notificaram no dia 28 de março os clubes e seus cartolas envolvidos na negociação , além_de neymar e seus representantes ,
apontando operações que , segundo a dis , contêm irregularidades .
apontar,name,identificar,seu,representante,",",operação,que,",",v,v,det,n,cm,n,rel,cm,0.182,0.799,no

% cand=1185 eloize apontar identificar
% 37 37 a_cada dia , 116 pessoas morrem vítimas de armas de fogo no brasil . em 2012 , data dos últimos dados disponíveis , foram 42.416 mortos . a
principal causa das mortes foram homicídios , motivo apontado em 95 % dos casos .
apontar,name,identificar,homicídio,",",motivo,em,95,"%",v,v,n,cm,n,pr,num,NC,0.182,0.799,no
```

Figura 1. Formato de entrada (*.arff*) para a ferramenta NEPaLE

Na Figura 2 encontra-se uma ilustração da interface de avaliação da NEPaLE. Depois de carregar o arquivo de avaliação, todos os pares de candidatas a serem avaliados, bem como os contextos de avaliação, são apresentados para o usuário, conforme mostrado na figura (as candidatas a paráfrases são destacadas na cor azul). No exemplo da figura as candidatas aparecem na sua forma canônica, pois o algoritmo de aprendizado sem fim trabalha com os textos lematizados. Vale dizer que o mesmo conjunto de dados pode ser analisado por vários avaliadores, porém para cada avaliador deve ser criado um arquivo *arff*[4].

Durante a avaliação, o avaliador verifica se determinado par de candidatas a paráfrases é intercambiável naquele contexto (ou seja, se ao substituir uma candidata pela outra o sentido permanece o mesmo), e seleciona a opção correspondente por meio dos botões *Sim* e *Não*. Caso ele fique em dúvida num primeiro momento, ele pode escolher a opção *Não Sei*, e fazer a sua escolha pelo Sim ou pelo Não mais tarde. Não é necessário avaliar todo o conjunto de dados de uma só vez. O avaliador pode interromper a avaliação a qualquer tempo e retornar posteriormente para finalizá-la. O arquivo *arff* só será alterado com a inclusão da resposta (*yes* ou *no*) no final da instância correspondente, quando o avaliador escolher a opção *Sim* ou a opção *Não* (como pode ser observado no *arff* da Figura 1).

[4] Esse arquivo não é construído pela NEPaLE. Em geral, ele é obtido como resultado do algoritmo de extração de paráfrases.

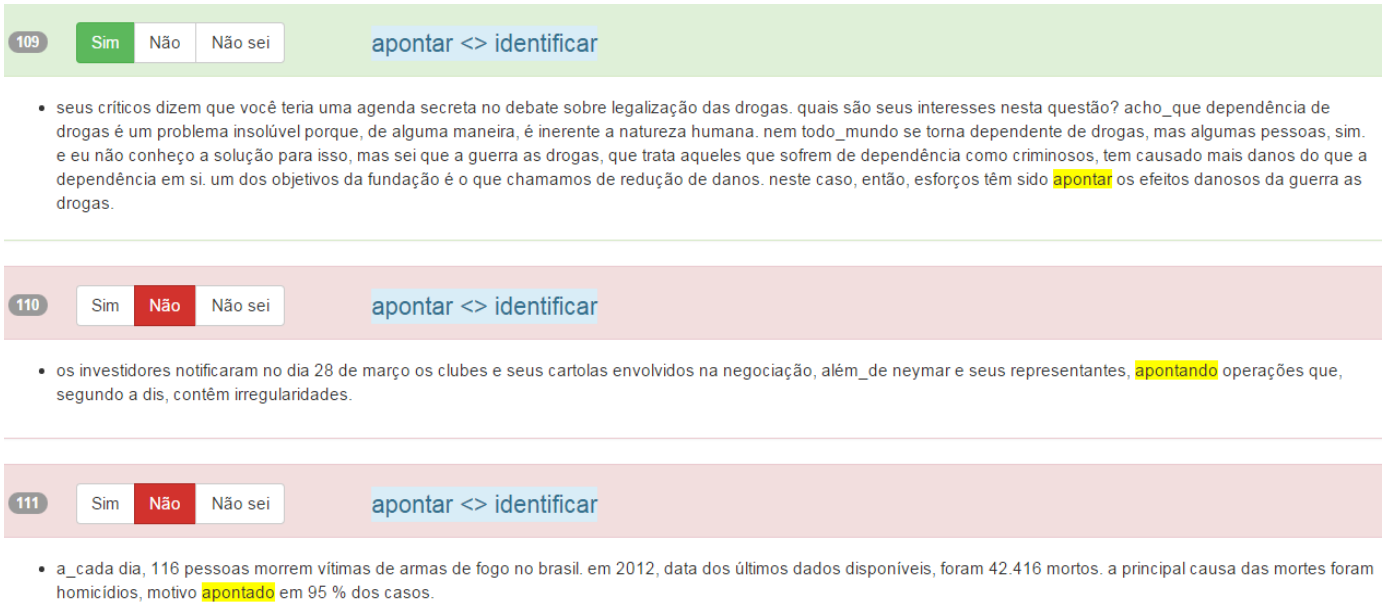

Figura 2: Interface da ferramenta NEPaLE

3. Conclusão

Este artigo apresentou a construção de uma ferramenta computacional para suportar a avaliação *online* de paráfrases, a NEPaLE. Trata-se de uma ferramenta independente de língua, que, além de possibilitar a avaliação de paráfrases, também pode ser usada como apoio para a anotação de paráfrases, sinônimos, etc., visando a construção de dados de treinamento e teste de algoritmos de aprendizado de máquina que utilizam como entrada arquivos no formato *arff*.

Encontram-se em fase de implementação algumas medidas estatísticas (como o coeficiente Kappa (Carletta, 1996), que possibilitam calcular a concordância entre os avaliadores na avaliação de um mesmo conjunto de dados. Outras funcionalidades também previstas são a geração de relatórios (ou gráficos), representando a proporção de exemplos julgados como Sim (isto é, que indicam paráfrases) e como Não (ou seja, que não indicam paráfrases), bem como o total de exemplos já analisados e o total de exemplos que ainda precisam ser avaliados.

Agradecimentos

Esse trabalho faz parte do projeto "Aprendendo com a *web* a traduzir e parafrasear textos", processo nº 2013/11811-0, da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). Vale ressaltar que as opiniões, hipóteses e conclusões ou recomendações expressas neste material são de responsabilidades dos autores e não necessariamente refletem a visão da FAPESP. Também agradecemos ao Instituto Federal de São Paulo – IFSP pelo apoio financeiro.

Referências

Androutsopoulos, I. and Malakasiotis, P. (2010). A Survey of Paraphrasing and Textual Entailment Methods. *Journal of Artificial Intelligence Research*, 38, p. 135-187.

Bannard, C. and Callison-Burch, C. (2005). Paraphrasing with Bilingual Parallel Corpora. In: *Proceedings of the 43rd Annual Meeting of Association for Computational Linguistics - ACL*, p. 597–604.

Barreiro, A. M. (2008). Make it Simple with Paraphrases: Automated Paraphrasing for Authoring Aids and Machine Translation. PhD dissertation. Faculdade de Letras da Universidade do Porto. Porto, Portugal.

Barzilay, R. and McKeown, K. (2001). Extracting Paraphrases from a Parallel Corpora. In: *Proceedings of Association for Computational Linguistics - ACL*, p. 50-57.

Barzilay, R. and McKeown, K. (2005). Sentence Fusion for Multi-document News Summarization. *Computational Linguistics*, Vol. 31, nº 3, pp. 297-327.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, v. 22, n. 2, pp. 249-254.

Duboue, P. A. and Chu-Carroll, J. (2006). Answering the Question You Wish They Had Asked: The Impact of Paraphrasing for Question Answering. In: *Proceedings of the Human Language Technology – HLT/NAACL*, p. 33-36.

Riezler, S.; Vasserman, A.; Tsochantaridis, I.; Mittal, V.; Liu, Y. (2007). Statistical Machine Translation for Query Expansion in Answer Retrieval. In: *Proceedings of the 45th Annual Meeting of Association for Computational Linguistics - ACL*, p. 464–471.

Witten, I.H. and Hall, M.A. (2005). Data Mining: Practical Machine Learning Tools and Techniques. 2ª ed., Elsevier: São Franscisco, CA, EUA.

Zhao, S. et al. (2009). Extracting Paraphrase Patterns from Bilingual Parallel Corpora. *Natural Language Engineering*, 15 (4), p. 503-526.