Extração de paráfrases em português a partir de léxicos bilíngues: um estudo de caso

Paulo César Polastri[1,2], Helena de Medeiros Caseli[1,2], Eloize Rossi Marques Seno[2,3]

[1] Departamento de Computação, [2] LALIC,
Universidade Federal de São Carlos (UFSCar), São Carlos/SP, Brasil
{paulo.polastri, helenacaseli}@dc.ufscar.br
[3] Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP),
São Carlos/SP, Brasil
eloize@ifsp.edu.br

Abstract. Este artigo apresenta um método de extração de paráfrases a partir de léxicos bilíngues baseando-se na ideia de que duas ou mais traduções de uma mesma palavra são consideradas paráfrases. O resultado da aplicação deste método é a geração de um léxico de paráfrases, em nível de palavras, do português do Brasil. Um cálculo probabilístico foi empregado para enfatizar os alinhamentos que mais ocorreram. Deste modo, além de resultados consistentes, a abrangência de informação é maior em relação a léxicos gerados a partir de corpus monolíngues.

Keywords: Léxico bilíngue, paráfrase, pares alinhados, português do Brasil.

1 Introdução

Paráfrases são formas alternativas de transmitir uma informação utilizando diferentes termos linguísticos [4]. Segundo [4], as paráfrases são classificadas em três níveis de granularidade. Isso pode observado nos exemplos da Tabela 1, onde, considerando o contexto no qual está inserido, o exemplo (1) representa um par de paráfrases com granularidade em nível de palavras (formado por palavras simples), o (2) apresenta um par de paráfrases em nível de sintagmas e o (3) apresenta um par de paráfrases em nível de sentenças.

Tabela 1. Exemplos de paráfrases.

(1)	escutar
	ouvir
(2)	Agulhas Negras
	Academia Militar Agulhas Negras
(3)	O clube contratou um novo jogador
	Um novo jogador foi contratado pelo clube

A identificação de paráfrases é uma das principais tarefas dentre os trabalhos que lidam com paráfrases disponíveis na literatura, como em [3, 4, 5]. As paráfrases são importantes para diversas áreas do Processamento de Língua Natural (PLN) como: na Sumarização Automática Multidocumento, onde a identificação de paráfrases auxilia na identificação de informações redundantes; na Tradução Automática, onde paráfrases podem ser usadas pala melhorar a qualidade da tradução; e na Geração de Língua Natural, onde as paráfrases são utilizadas para a criação de textos mais fluentes e para aumentar a variedade textual.

Neste trabalho é apresentado um estudo de caso sobre um método de extração de paráfrases em português, em nível de palavras, através de léxicos bilíngues, utilizando um idioma como pivô. Com a aplicação deste método espera-se obter um léxico de paráfrases em português que possa ser utilizado pela comunidade científica em diversas aplicações do PLN.

A sequência deste trabalho está organizada da seguinte forma. Na seção 2 é abordado com mais detalhes o método utilizado para extração de paráfrases a partir de léxicos bilíngues. Na seção 3 são apresentados os experimentos preliminares e, na seção 4, os resultados juntamente com uma comparação com o trabalho de [3], método que serviu de base para a proposta aqui apresentada. Por fim, na seção 5 são apresentadas as conclusões e também alguns trabalhos futuros.

2 Extração de paráfrases em português

O objetivo deste trabalho é a extração de paráfrases em português a partir de léxicos bilíngues, utilizando um idioma como pivô. O método de extração de paráfrases utilizado é baseado no trabalho apresentado por [3]. Esse método parte do pressuposto de que uma frase X em um idioma A (alvo) pode ter várias traduções no idioma B (pivô) e essas podem ser traduzidas para o idioma A gerando frases Y correspondentes (paráfrases) à frase X.

O método empregado por [3] consiste em identificar alinhamentos entre itens lexicais do mesmo idioma (inglês, no caso) através de traduções em um idioma pivô (alemão, no caso). A partir da identificação dos alinhamentos, a probabilidade dos itens se parafrasearem é calculada com base no número de ocorrências, ou seja, através da divisão do número de alinhamentos que determinado par obteve pela quantidade total de ocorrências de alinhamentos em que a palavra alvo aparece. Mais detalhes do método podem ser obtidos em [3].

Os léxicos bilíngues utilizados no estudo de caso descrito neste artigo foram gerados por [1] através do alinhamento de sentenças do Corpus FAPESP [7] – conjuntos de textos da revista científica Pesquisa FAPESP originalmente escritos em português do Brasil e traduzidos para inglês e o espanhol – utilizando o alinhador lexical GIZA++ [2]. Dos léxicos gerados por [1], foram utilizados neste trabalho o léxico português-inglês, contendo 159.814 pares de alinhamentos e o léxico inglês-português, contendo 158.037 pares de alinhamentos.

Ambos os léxicos foram submetidos a pré-processamentos automáticos que incluíram a lematização e a remoção de *stopwords*. A lematização (processo de redução de palavras para seu lema) foi realizada com o intuito de agrupar todas as flexões de uma palavra em um único item, visando evitar a baixa frequência dos alinhamentos. A lematização foi realizada usando o LematizadorV1[1]. Outro processo realizado foi a exclusão de *stopwords* (palavras que ocorrem com grande frequência em determinado idioma). Neste processo foi utilizada uma lista de *stopwords* contento palavras de classes fechadas, como preposições, conjunções e artigos, em português. Esse processo foi realizado com a intenção de evitar alinhamentos sem importância para a pesquisa. Após a execução desses dois processos, os léxicos resultantes foram utilizados neste trabalho conforme descrito na próxima seção.

3 Experimentos

A partir do método proposto em [3], os léxicos bilíngues resultantes dos processos de lematização e exclusão de *stopwords* (vide seção 2) foram alinhados, visando obter pares de paráfrases em português, utilizando o inglês como idioma pivô. A partir da aplicação do método, foram gerados 824.928 pares alinhados. Aleatoriamente, 20 pares de alinhamentos foram selecionados para serem avaliados. Para a seleção dessa amostra de avaliação, alinhamentos com probabilidade de paráfrase inferior a 10% foram desconsiderados.

Cada palavra do par de alinhamentos foi avaliada em sentenças originais em que ocorreram no Corpus FAPESP. Para tanto, foram utilizadas 100 sentenças (5 sentenças para cada par). O objetivo da avaliação era verificar se os itens do par alinhado pelo método poderiam ser intercambiados em qualquer sentença em que um dos itens do par ocorreu. Em cada sentença, a ocorrência de qualquer um dos itens pertencentes ao par avaliado foi substituída por lacunas. A Tabela 2 traz exemplos de pares de palavras candidatas a paráfrase extraídas pelo método.

Tabela 2. Exemplos de pares de palavras candidatas a paráfrase extraídas.

Par candidato	***Sentença de avaliação***
Asfaltamento/pavimentação	Estão até financiando um uma parte do _____ rodovia Cuiabá–Santarém
Asfaltamento/pavimentação	Sem ____, só serve para passar gado.
Muralha/parede	A segurança de Israel teve benefícios ao seguir a lógica da ___ de Ferro.

1 Disponível em: http://www.icmc.usp.br/pessoas/taspardo/sucinto/resources.html. Acesso em: 22/05/2014.

Para avaliar os alinhamentos, três juízes falantes nativos do português do Brasil foram instruídos a substituir a lacuna pelas palavras do par, de forma a atribuir a cada nova sentença gerada uma pontuação de 0 a 2, sendo 0 para sentenças incorretas ou sem sentido, 1 para sentenças compreensíveis mas com leve alteração no sentido e 2 para sentenças perfeitas, considerando utilizar qualquer item do par avaliado.

4 Resultados

As avaliações foram feitas por três juízes falantes nativos do português do Brasil, como descrito na seção 3. Cada juiz analisou 100 sentenças e as porcentagens de paráfrases avaliadas como corretas, por cada juiz, bem como a média dos três juízes são detalhadas na Tabela 3:

Tabela 3. Resultados da avaliação.

Avaliadores	**Juiz1**	**Juiz2**	**Juiz3**	**Média**
Paráfrases corretas	38%	42%	52%	44%

Além dos resultados acima, a concordância entre os julgamentos foi calculada através da medida Kappa [6] apresentado o valor de concordância de κ=0,382. Isso se deve ao fato que alinhamentos com probabilidade baixa (como 11,2%) foram incluídos na avaliação. Futuramente, o experimento será avaliado considerando apenas itens com maior grau de confiabilidade, com isso espera-se um resultado melhor em relação à concordância entre avaliadores.

Entre os alinhamentos avaliados é possível notar tanto resultados considerados pelos juízes como corretos, como é o caso dos alinhamentos entre asfaltamento/pavimentação, pesticida/praguicida, resultados parcialmente corretos, como é o caso dos pares endurecimento/solidificação e passeio/viagem, e também resultados ruins e não esperados, como é o caso dos pares paixão/passional e enrolar/corda. Esses alinhamentos inesperados serão melhor avaliados em trabalhos futuros e não se descarta o uso de filtros para eliminar candidatos, por exemplo, com categorias gramaticais distintas como é o caso de enrolar/corda.

Em comparação com o trabalho de [3], considerando apenas alinhamentos automáticos, assim como foi feito no presente trabalho, os resultados são bastante semelhantes. Entre as sentenças avaliadas, em média 48,9% foram consideradas corretas pelos avaliadores, uma diferença de apenas 4,9% do método aqui apresentado (44%). Contudo, a avaliação da medida de concordância Kappa apresentou uma diferença maior: κ=0,605.

5 Conclusão e trabalhos futuros

Esse trabalho utilizou um método de geração de um léxico monolíngue a partir de dois léxicos bilíngues, mostrando que paráfrases podem ser identificadas a partir de traduções em uma língua pivô. Além disso, para cada alinhamento foi calculada a probabilidade de ocorrência, de acordo com o método descrito acima.

A ferramenta criada neste trabalho possibilita a geração de um léxico monolíngue em português a partir de dois léxicos bilíngues, podendo ser utilizado qualquer idioma como pivô. O resultado desse processo é uma lista de paráfrases em português do Brasil que poderá ser utilizada em várias aplicações do PLN.

Para lidar com casos em que candidatos envolvendo termos que diferentes *part-of-speech* (como corda/amarrar, por exemplo) foram gerados, será analisada a possibilidade de utilização de filtros. Contudo, vale lembrar que a estratégia adotada nos experimentos aqui descritos utiliza a tradução como meio para gerar os candidatos, e mudanças de *part-of-speech* são frequentes em textos traduzidos.

Além disso, outros testes deverão ser feitos considerando diferentes limiares de probabilidade de paráfrase. Neste experimento, alinhamentos com probabilidade de paráfrase inferior a 10% foram desconsiderados. Em experimentos futuros, poderão ser testados os alinhamentos com percentagem de probabilidade diferentes, como apenas probabilidades maiores que 30% ou 50%, visando atingir melhores resultados e, possivelmente, maior concordância entre os avaliadores.

Por fim, vale mencionar que já está em andamento a extensão dos experimentos aqui apresentados usando, além do inglês, também o espanhol como língua pivô. Outra extensão também em andamento, é a aplicação deste método para extração de paráfrases em nível de sintagmas.

Agradecimentos

Este trabalho é parte dos processos no. 2013/11811-0 e 2013/50757-0 (AIM-WEST), Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), à qual agradecemos.

Referências

1. Caseli, H. Indução de léxicos bilíngues e regras para a tradução automática. Tese (Ph.D) – Instituto de Ciências Matemáticas e de Computação, (2007)
2. Och, F.J., Ney, H. “Improved statistical alignment models”. In Proceedings of the 38th Annual Meeting of the ACL. Hong Kong, China, pp. 440-447, (2000)
3. Bannard, C., Callison-Burch, C. “Paraphrasing with bilingual parallel corpora”. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), Ann Arbor, Michigan, USA, pp. 597-604 (2005)
4. Barzilay, R., McKeown, K. R. “Extracting paraphrases from a parallel corpus”. In Proceedings of ACL/EACL, Toulouse, pp. 50-57 (2001)
5. Pang, B., Knight, C., Marcu, D. “Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences”. In Proceedings of HLT/NAACL, Edmonton, Canada, pp. 102-109 (2003)

6. Carletta, J. “Assessing agreement on classification tasks: The kappa statistic”. Computational Linguistics, Eindhoven, Netherlands, pp. 249-254, (1996)
7. Aziz, W., Specia, L. “Fully Automatic Compilation of a Portuguese-English Parallel Corpus for Statistical Machine Translation”. In Proceedings of STIL 2011, Cuiabá, Brazil, pp. 234-238, (2011)