Aprendizado de Máquina Sem-Fim para Indução Automática de Léxico Bilíngue

Thiago Lima Vieira and Helena de Medeiros Caseli

Departamento de Computação, Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 – 13.565-905, São Carlos, São Paulo, Brasil
{thiago.lima.vieira,helenacaseli}@dc.ufscar.br
http://www.lalic.dc.ufscar.br

Resumo Este trabalho apresenta os resultados da aplicação do Aprendizado de Máquina Sem-Fim (AMSF) para a indução de um léxico bilíngue Inglês-Português. Baseado no processo humano de aprendizado, o AMSF é uma estratégia de aprendizado contínuo, que utiliza os conhecimentos já adquiridos para aprender novas informações e, assim, melhorar seu desempenho. O AMSF se mostra interessante neste contexto pois utiliza a Internet como fonte de conhecimento e combina métodos diferentes de extração para melhorar o resultado final. Em resultados preliminares de avaliação, apresentados neste artigo, constatou-se uma precisão de 65% no léxico bilíngue resultante.

Keywords: léxico bilíngue, aprendizado de máquina sem-fim

1 Introdução

Um Léxico Bilíngue pode ser considerado um conjunto de itens lexicais (palavra ou unidades multipalavras) de uma língua fonte acompanhados de seus respectivos equivalentes (tradução) na língua alvo. Por exemplo, *"bread"* (inglês) com "pão" (português), ou *"throw away"* (inglês) com "jogar fora" (português). Quando o léxico é extraído automaticamente de um corpus paralelo[1] é comum atribuir à cada entrada a probabilidade ou frequência de ocorrência no *corpus* para indicar a confiança do par.

O léxico bilíngue é um recurso muito importante em aplicações multilíngue, como principal exemplo a Tradução Automática (TA). Na TA estatística, por exemplo, as entradas do léxico bilíngue podem ser inseridas nos modelos de língua e tradução para enriquecê-los, aumentando a cobertura e, assim, melhorando o resultado final da TA. Nos levantamentos realizados por [4] e [14] constatou-se que a maior percentagem de erros encontrados na TA inglês-português decorrem de erros léxicos. Desse modo, a indução automática, constante e incremental de léxicos bilíngues se faz significativa para melhorar o resultado final da TA e de outras aplicações multilíngues.

[1] *Corpus* paralelo é um conjunto de pares de textos onde o texto na língua fonte é acompanhado do seu respectivo texto traduzido na língua alvo.

Um léxico bilíngue pode ser construído manual ou automaticamente. A construção manual, embora seja mais precisa, demanda tempo e tem alto custo de implementação. Como alternativa, os pares de tradução que dão origem ao léxico podem ser aprendidos automaticamente a partir de *corpus* paralelo. Nesse contexto, a estratégia de Aprendizado de Máquina Sem-Fim (AMSF) surge como uma opção.

O AMSF é uma estratégia de aprendizado de máquina semi-supervisionado que tenta reproduzir o aprendizado humano: primeiramente, fatos simples são aprendidos e, a cada dia, esse conhecimento é utilizado para aprender novos fatos mais complexos no futuro [17]. Sistemas de AMSF executam constantemente e usam como fonte de conhecimento (dados de treinamento) a Internet devido, principalmente, ao fluxo crescente de novos conteúdos que são inseridos a todo momento [6].

À vista disso, a utilização de AMSF para indução automática de léxico bilíngue se faz interessante, pois o léxico bilíngue resultante tem características desejáveis para um recurso de Processamento de Língua Natural: (i) é mais abrangente, devido à utilização de uma fonte dinâmica (a Internet) de conhecimento, (ii) as variações linguísticas podem ser identificadas pela execução constante, mesmo que não sejam frequentes o bastante para serem consideradas pelos métodos tradicionais baseados em corpus e (iii) a geração do léxico ocorre praticamente na mesma velocidade do surgimento das variações de uso da língua.

Este artigo está organizado como segue. A próxima seção apresenta alguns trabalhos relacionados a este. Em seguida, é apresentada uma visão geral sobre a arquitetura e funcionamento do aprendiz sem-fim utilizado para indução automática do léxico bilíngue, seguida da seção que descreve os resultados obtidos no experimento realizado. Por fim, a seção com algumas conclusões e trabalhos futuros.

2 Trabalhos relacionados

Os trabalhos de extração de léxico podem ser divididos, de forma geral, pelo método: manual ou automático. A construção manual resulta em um léxico rico e minucioso da língua, porém, essa abordagem é cara, pois exige especialistas e tempo, além de tornar o crescimento do léxico algo complexo. Por outro lado, a maneira automática permite a construção do léxico de forma rápida, barata e altamente escalável, contudo, com menor precisão.

A WordNet de Princeton é o principal trabalho relacionado à construção manual de léxico. Criado para emular o léxico mental, a WordNet é uma grande base de dados lexical para o inglês americano, composta por substantivos, verbos, adjetivos e advérbios agrupados em conjuntos de sinônimos [18].

O projeto da WordNet de Princeton foi muito bem aceito pela comunidade de PLN, e diversos trabalhos para criação de *wordnets* mono e multi-idiomas surgiram [22,13,8].

No entanto, a construção manual de léxico é uma tarefa que necessita de especialistas para sua elaboração, é demorada e cara, neste contexto a indução

automática surge com uma alternativa barata e rápida, embora menos precisa. A abordagem automática baseia-se em modelos estatísticos para induzir o léxico automaticamente a partir de *corpora*.

Segundo [15], a maioria dos trabalhos que utilizam métodos estatísticos para construção de léxico são variações do algoritmo guloso, que pode ser resumido em quatro passos:

1. Primeiro escolhe-se a função de similaridade utilizada para calcular a distância entre as palavras da língua fonte (LF) com as palavras da língua alvo (LA);
2. Calcula-se a pontuação de associação entre um conjunto de palavras da LF e LA;
3. Ordena-se a lista de pares de palavras com relação à pontuação de associação;
4. Finalmente, escolhe-se um ponto de corte, o qual é utilizado para eliminar os pares com pontuação inferior.

Este esquema, com algumas variações, é aplicado na maioria dos trabalhos de extração automática da literatura [2,11,23,9,16,10,12,5,7,20], assim como neste artigo.

Observando estes trabalhos [23,16,12,5] fica evidente a necessidade de uma etapa de pré-processamento dos textos de entrada antes da fase de extração do léxico. De modo geral, esse pré-processamento envolve alguma forma de alinhamento entre textos paralelos para explorar a semelhança entre a posição das palavras que ocorrem próximas no texto fonte e alvo.

Os trabalhos [10,12] tiram proveito da similaridade entre as línguas, calculando a distância de edição entre os itens lexicais. Já os trabalhos [11,23] utilizam de informações estatísticas como informação mútua, coocorrência e verossimilhança.

Há várias estratégias que podem ser utilizadas para a identificação de equivalentes lexicais bilíngues. No entanto, todas as abordagens da literatura utilizam *corpus* de tamanho fixo e um grupo pequeno de medidas baseadas nessa característica. O aprendiz incremental, proposto neste artigo, tem o *corpus* de entrada de tamanho variável a cada iteração e aposta na combinação de diferentes medidas de similaridades para obter um resultador melhor.

3 Never-Ending Bilingual Equivalent Learner

Para a indução automática de léxico bilíngue através do AMSF foi desenvolvido um sistema aprendiz batizado de *Never-Ending Bilingual Equivalent Learner* (NEBEL). A tarefa do NEBEL é vasculhar a Internet em busca de textos paralelos, processar esses textos para extrair deles possíveis candidatos à tradução e promover alguns desses candidatos considerando-os, de fato, como traduções. No NEBEL, os candidatos à tradução são as *possibilidades* e os candidatos promovidos são as *crenças*.

A figura 1 apresenta a arquitetura do NEBEL, a qual é dividida em quatro módulos principais explicados em detalhes a seguir: Coletor, Pré-processador, Processador e Promotor.

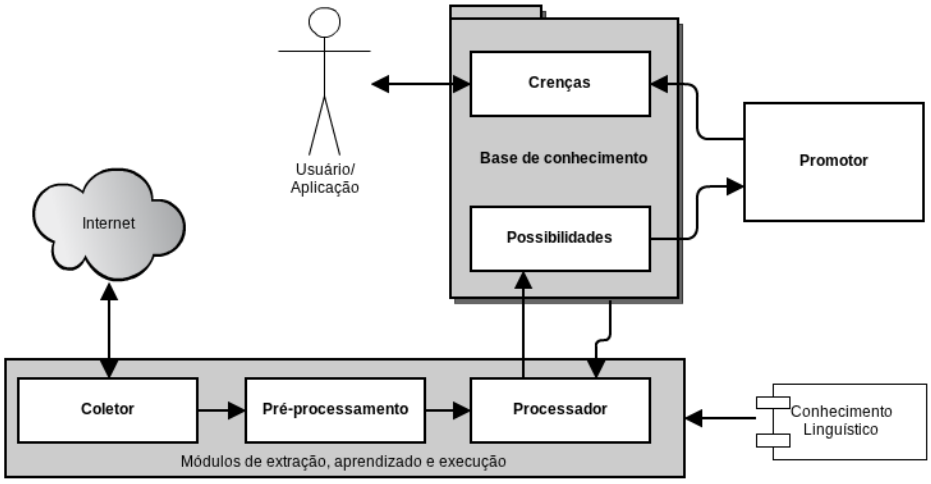

Figura 1. Arquitetura básica no NEBEL.

Essa aplicação é baseada no *Never-Ending Language Learner* (NELL)[2], o primeiro sistema de AMSF desenvolvido pelo projeto *Read The Web*[17].

O foco do NELL e do NEBEL são distintos. O NELL lê a Internet para extrair categorias (fatos) e relações entre as categorias na língua inglesa, por exemplo: *is_a(shakespeare, writer)*; *writer_wrote_book(shakespeare, hamlet)*. O NEBEL possui apenas uma relação (é_tradução) e as categorias aprendidas são os equivalente lexicais para o par de línguas inglês-português, por exemplo: é_tradução(*"love"*, "amor"); é_tradução(*"let it go"*, "deixe para lá"). Assim sendo, os dados de entrada para treinamento, os métodos de extração, e o conhecimento resultante são diferentes.

A intersecção dos trabalhos se situa na aplicação da estratégia de AMSF: (i) utilização da Internet como fonte de conhecimento, (ii) combinação de diferentes métodos de extração, e (iii) aplicação do conhecimento já aprendido para a melhoria da capacidade de extração de novos conhecimentos.

A estratégia de indução automática de léxico desenvolvida atualmente no NEBEL necessita de corpus paralelo. Na Internet é possível encontrar esse tipo de corpus paralelo, por exemplo, em forma de letras de músicas e legendas de filmes ou séries.[3].

O módulo Coletor é responsável pelo acesso e coleta dos textos nos respectivos repositórios (de legendas e de letras de músicas). Para tanto, foram desenvolvidos *crawlers* específicos. O *crawler* de letras de música, por exemplo, ao acessar uma determinada música recebe informações sobre a tradução e músicas relacionadas,

[2] NELL: `http://rtw.ml.cmu.edu/rtw/`

[3] Dois repositórios se destacam pela sua facilidade de uso: OpenSubtitles [`http://www.opensubtitles.org/`] e Vagalume [`http://www.vagalume.com.br/`]

que ele busca em seguida. Assim, o módulo Coletor extrai os textos da Internet para formar o corpus paralelo.

Para lidar com as particularidades de cada uma das fontes de conhecimento usadas pelo NEBEL, há uma etapa de pré-processamento realizada pelo módulo Pré-processador. O pré-processamento envolve a limpeza do texto, *tokenização*, etiquetação morfossintática[4], alinhamento automático sentencial[5] e lexical[6]. Informações detalhadas sobre todas as ferramentas utilizadas neste módulo podem ser encontradas no Portal de Tradução Automática PorTAl [21].[7]

O módulo Processador é o coração do sistema. Ele utiliza o conhecimento já adquirido (crenças), armazenado na base de conhecimento, para extrair candidatos através de diferentes métodos de extração de léxico bilíngue. Para tanto, cada iteração do Processador tem como entrada um par de sentenças paralelas já pré-processadas. Como é possível que qualquer *token* da sentença fonte seja tradução de qualquer *token* da sentença alvo, assim como, qualquer combinação de *tokens* vizinhos na sentença fonte seja tradução de qualquer combinação de *tokens* vizinhos na sentença alvo, são gerados candidatos para todas as possibilidades, e medidas de similaridades são aplicadas para estipular quão provável a possibilidade é de ser, de fato, uma crença.

O NEBEL atualmente conta com 23 medidas de similaridades divididas em quatro grupos:

1. **Medidas Simples:** exploram características de fácil observação e constatação entre os itens lexicais. Ex.: frequência do item lexical fonte, coocorrência do par, tamanho de *tokens*, tamanho de caracteres, posição na sentença e outras.
2. **Medidas Linguísticas:** se baseiam em alguma observação mais profunda da língua. Seja a relação entre os idiomas ou em características da língua que podem indicar uma tendência na relação de tradução. Ex.: sinônimos, antônimos, padrões de prefixo e sufixo, e categoria gramatical.
3. **Medidas Baseadas em Distância de Edição:** aproveitam das semelhanças entre palavras, e consecutivamente itens lexicais, para capturar trechos de caracteres semelhantes em ambas as línguas. Ex.: distância de Levenshtein, distância de Damerau Levenshtein, coeficiente Dice, Jaro Winkler, e Longest Common Sequence.
4. **Medidas Estatísticas:** utilizam de cálculos mais elaborados, que envolvem o corpus total na predição. Ex.: verossimilhança e informação mútua.

Com o suporte destas medidas, o módulo Processador gera um aglomerado de possibilidades que são analisadas posteriormente pelo Promotor. O Promotor combina o resultado de todas as medidas por meio de um algoritmo de aprendizado de máquina e classifica as possibilidades como tradução (promovendo-os

[4] O etiquetador morfossintático utilizado pelo NEBEL é o do *toolkit* de tradutução automática de código-aberto Apertium, com os dados linguísticos de [4].
[5] O alinhamento sentencial foi realizado pela ferramenta TCAlign [3].
[6] O alinhamento lexical foi realizado pela ferramenta GIZA++ [19].
[7] Disponível em: `http://www.lalic.dc.ufscar.br/portal`. Acesso em: 22 set. 2014.

à crença) ou não. Atualmente, o módulo Promotor do NEBEL utiliza a implementação do classificador *Naive Bayes* fornecida pelo Weka[8].

Para promover os primeiros candidatos a crença (S1), o classificador é treinado com o um léxico bilíngue semente (S0). As novas crenças mais as sementes ($S1 + S0$) são utilizadas para retreinar o classificador e promover os candidatos na segunda iteração (S2), o retreino do classificador é realizado sucessivamente ($S_n + ... + S2 + S1 + S0$) para melhorar sua capacidade de aprendizado. No entanto, já é conhecido que essa abordagem pode acarretar em acumulo de erro e consecutivamente desvio de conceitos e *overfitting*, por isso, a supervisão humana de tempos em tempos é necessária.

4 Experimento e Resultados

O experimento aqui apresentado é preliminar e foi realizado considerando-se a execução do NEBEL com apenas uma passada em um conjunto de 5989 mil pares de textos paralelos extraídos automaticamente dos repositórios de legendas e letras de músicas. Destes textos foram aprendidos 1682 mil pares que equivalentes lexicais (crenças).

O léxico bilíngue utilizado como semente (base de conhecimento inicial) na extração das possibilidades e no treinamento do Promotor foi induzido automaticamente por [5] a partir do *corpus* da Revista Pesquisa FAPESP[9] [1].

Os léxicos originais contêm 158.037 entradas inglês-português e 159.814 português-inglês. Todavia, com intuito de melhorar a precisão das entradas usadas como semente para o aprendiz, as entradas dos léxicos originais foram filtradas para conter apenas aquelas com frequência maior do que cem, gerando os léxicos sementes de tamanho reduzido: 3.382 inglês-português e 3.573 português-inglês.

Mil entradas do léxico foram avaliadas manualmente e obteve-se uma precisão de 65%.[10] Nessa avaliação não foi possível calcular a cobertura do léxico resultante uma vez que como os textos de entrada vêm da Internet, não se sabe a quantidade total de equivalentes lexicais que poderiam ser obtidos do corpus de entrada.

Dez exemplos de pares bilíngues, extraídos aleatoriamente, são apresentados na tabela 1.

5 Considerações Finais

O AMSF para indução automática de léxico bilíngue tem mostrado resultados satisfatórios, mesmo com a exigência de capacidade de processamento elevado.

[8] Disponível em: `http://www.cs.waikato.ac.nz/ml/weka/`. Acesso em: 22 set. 2014

[9] Revista Pesquisa FAPESP: `http://revistapesquisa.fapesp.br`. *Corpus* disponível para download em: `http://www.nilc.icmc.usp.br/nilc/tools/Fapesp\%20Corpora.htm`

[10] O avaliador possuía domínio de ambas as línguas e podia consultar o contexto do qual o par foi extraído.

Tabela 1. Dez pares de equivalentes lexicais extraídos pelo NEBEL.

Inglês	Português
"hallelujah"	"aleluia"
"i feel"	"me sinto"
"right_now"	"agora"
"i don't wanna"	"eu não quero"
"i don't know"	"eu não sei"
"man"	"cara"
"'s"	"isso"
"you"	"lhe"
"cuz"	"porque"
"whatcha whatcha"	"o que"

Vale ressaltar que o NEBEL é independente de língua, mas possui dependência das ferramentas e recursos externos como: etiquetador, léxico bilíngue semente, *corpus* paralelo de entrada, léxico de sinônimos e antônimos. Consequentemente a adaptação do aprendiz para outras línguas depende desses recursos e ferramentas, assim como a qualidade do léxico bilíngue resultante.

Os autores acreditam que o AMSF é uma poderosa alternativa para a geração de recursos linguísticos necessários no Processamento de Língua Natural, não apenas recursos bilíngues, mas, por exemplo, na indução automática de paráfrases, extração de relações semânticas e outros.

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. Os autores deste artigo também agradecem à CAPES/CNPq pelo apoio financeiro.

Referências

1. Aziz, W. & Specia, L.: Fully Automatic Compilation of a Portuguese-English Parallel Corpus for Statistical Machine Translation. STIL 2011, (2011).
2. Brown, P. F.; Cocke, J.; Pietra, S. A. D.; Pietra, V. J. D.; Jelinek, F.; Lafferty, J. D.; Mercer, R. L. & Roossin, P. S.: A statistical approach to machine translation. Computational linguistics, MIT Press, 1990, 16, 79-85.
3. Caseli, H.: Alinhamento sentencial de textos paralelos português-inglês. Uuniversidade de São Paulo, (2003).
4. Caseli, H. M.; Nunes, M. G. V. & Forcada, M. L.: Automatic induction of bilingual resources from aligned parallel corpora: application to shallow-transfer machine translation. Machine Translation, (2006), 20:4, 227-245.
5. Caseli, H. d. M.: Indução de léxicos bilíngües e regras para a tradução automática. Biblioteca Digital de Teses e Dissertações da USP, (2007).

6. Carlson, A.; Betteridge, J.; Kisiel, B.; Settles, B.; Jr., E. R. H. & Mitchell, T. M.: Toward an Architecture for Never-Ending Language Learning. In: Proceedings of the Twenty-Fourth Conference on Artificial Intelligence, (2010).
7. Di-Felippo, A. & de Barcellos Almeida, G. M.: Uma metodologia para o desenvolvimento de wordnets terminológicas em português do Brasil. TradTerm, (2010), 16, 365-395.
8. Dias-Da-Silva, B.; Di-Felippo, A. & Hasegawa, R.: Methods and tools for encoding the wordnet. br sentences, concept glosses, and conceptual-semantic relations. Computational Processing of the Portuguese Language, Springer, (2006), 120-130.
9. Fung, P.: A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. Proceedings of the 33rd annual meeting on Association for Computational Linguistics, (1995), 236-243.
10. Fung, P. & Yee, L. Y.: An IR approach for translating new words from nonparallel, comparable texts. Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1, Association for Computational Linguistics, (1998), 414-420.
11. Gale, W. A. & Church, K. W.: Identifying word correspondences in parallel texts. Proceedings of the workshop on Speech and Natural Language, (1991), 152-157.
12. Koehn, P. & Knight, K.: Learning a translation lexicon from monolingual corpora. Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition-Volume 9, (2002), 9-16.
13. Marrafa, P.; Amaro, R.; Chaves, R. P.; Lourosa, S.; Martins, C. & Mendes, S.: WordNet.PT – Uma rede léxico-conceptual do Português on-line. XXI Encontro da Associação Portuguesa de Linguística, Porto, Portugal, (2005).
14. Martins, D. B. J.: Pós-edição automática de textos traduzidos automaticamente de inglês para português do Brasil. Centro de Ciências Exatas e de Tecnologia – Programa de Pós-graduação em Ciência da Computação, Universidade de São Carlos, (2014), 97.
15. Melamed, I. D.: Automatic construction of clean broad-coverage translation lexicons, (1996).
16. Melamed, I. D.: A scalable architecture for bilingual lexicography, (1997).
17. Mitchell, T. M.; Betteridge, J.; Carlson, A.; Hong, S. A.; Hruscka, E. a. L.-M. E. & Wang, S.: Never-Ending Language Learning: The ReadTheWeb Manifesto, (2008).
18. Miller, G. A.; Beckwith, R.; Fellbaum, C.; Gross, D. & Miller, K. J.: Introduction to wordnet: An on-line lexical database International journal of lexicography, Oxford Univ Press, (1990), 3, 235-244.
19. Och, F. J. & Ney, H.: A Systematic Comparison of Various Statistical Alignment Models. Association for Computational Linguistics, (2003), 29, 19-51.
20. Riesa, J. & Marcu, D.: Automatic Parallel Fragment Extraction from Noisy Data. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT), (2012).
21. Vieira, T. L. & Caseli, H. M.: PorTAl: Recursos e Ferramentas de Tradução Automáticapara o Português do Brasil. In: Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology (STIL), (2011), 179-183.
22. Vossen, P. & others: EuroWordNet: a multilingual database for information retrieval. Proceedings of the DELOS workshop on Cross-language Information Retrieval, (1997), 5-7.
23. Wu, D. & Xia, X.: Learning an English-Chinese lexicon from a parallel corpus. Proceedings of the First Conference of the Association for Machine Translation in the Americas, (1994), 206-213.