http://dx.doi.org/10.1590/0102-445034728151307539

D E L T A

Uso de corpora comparáveis para filtrar dicionários bilíngues gerados por transitividade

Using comparable corpora to filter bilingual dictionaries generated by transitivity

Pablo Gamallo
(CITIUS - Universidade de Santiago de Compostela)

RESUMO

Este artigo propõe um método para a construção de novos dicionários bilíngues a partir de dicionários já existentes e da exploração de corpora comparáveis. Mais concretamente, um novo dicionário para um par de línguas é gerado em duas etapas: primeiro, cruzam-se dicionários bilíngues entre essas línguas e uma terceira intermediária e, segundo, o resultado do cruzamento, que contém um número elevado de traduções espúrias causadas pela ambiguidade das palavras da língua intermediária, filtra-se com apoio em textos de temática comparável nas duas línguas alvo. A qualidade do dicionário derivado é muito alta, próxima dos dicionários construídos manualmente. Descreveremos um caso de estudo onde criaremos um novo dicionário Inglês-Português com mais de 7.000 entradas bilíngues geradas pelo nosso método.

Palavras-chave: *processamento da língua natural; extração de informação; corpora comparáveis; dicionários bilíngues.*

ABSTRACT

This article proposes a method for building new bilingual dictionaries from existing ones and the use of comparable corpora. More precisely, a new bilingual dictionary with pairs in two target languages is built in two steps. First, a noisy dictionary is generated by transitivity by crossing two existing dictionaries containing translation pairs in one of the two target languages and an intermediary one. The result of crossing the two existing dictionaries gives rise to a noisy resource because of the ambiguity of words in the intermediary language. Second, odd translation pairs are filtered out by making use of a set of bilingual lexicons automatically extracted from comparable corpora. The quality of the filtered dictionary is very high, close to that of those dictionaries built by lexicographs. We also report a case study where a new, non noisy, English-Portuguese dictionary with more than 7,000 bilingual entries was automatically generated.

Key-words: *natural language processing; information extraction; comparable corpora; bilingual dictionaries.*

1. Introdução

A tradução automática baseada em regras precisa de dicionários bilíngues de ampla cobertura para oferecer traduções de qualidade. Nos últimos anos, o sistema de tradução automática Opentrad, de código aberto e multilíngue, tem alargado a sua oferta de pares de línguas sobre os que trabalha até atingir mais de 20 pares. No entanto, entre as línguas consideradas, existem ainda pares não explorados, nomeadamente Inglês-Português. De facto, a taxa de crescimento do número de dicionários bilíngues que requer um sistema de tradução multilíngue é uma função quadrática do número de línguas que o sistema traduz (Wherli et al., 2009). Ajudar a automatizar o processo de construção de novos dicionários é, portanto, uma tarefa crucial para reduzir drasticamente a quantidade de trabalho.

A estratégia mais natural para criar o novo dicionário é aproveitar a informação contida nos dicionários já existentes. Da mesma maneira que um lexicógrafo precisa dominar técnicas de trabalho para aproveitar o que já foi feito, o objectivo deste artigo é apresentar um método au-

tomático de elaboração de dicionários bilíngues a partir do cruzamento entre dicionários de pares de línguas já inseridos no sistema Opentrad, filtrando o resultado do cruzamento com informação extraída de corpora comparáveis. O método proposto é totalmente não supervisado e consiste, mais concretamente, nestas duas tarefas:

- Dados dois dicionários bilíngues para dois pares de línguas (*A*,*B*) e (*B*,*C*), geramos um novo par (*A*,*C*) por transitividade, onde *B* é a língua intermediária ou pivô. Por exemplo, a partir dos dicionários *(Inglês, Espanhol)* e *(Espanhol, Português), derivamos por transitividade um novo dicionário (Inglês, Português).*

- As correspondências bilíngues geradas a partir de termos ambíguos da língua intermediária são validadas e filtradas com base na similaridade distribucional computada automaticamente em corpora comparáveis, constituídos por textos nas duas línguas alvo: *A* e *C*. É dizer, precisamos de textos em Inglês e Português para corrigir os erros do novo dicionário construído por transitividade. Para podermos calcular a similaridade distribucional dos pares bilingues, os corpora comparáveis são analisados sintaticamente mediante dependências.

Neste artigo, apresentaremos um caso de estudo no que se descreve o processo de criação por transitividade de um novo dicionário *(Inglês, Português)* a partir de dois existentes, *(Inglês, Espanhol)* e *(Espanhol, Português)*, e filtrado com ajuda de um corpus comparável Inglês-Português. A língua intermediária é o Espanhol. Por exemplo, suponhamos que possuímos dois dicionários de verbos, *(Inglês, Espanhol)* e *(Espanhol, Português)*, com as seguintes correspondências bilíngues:

Tabela 1: Exemplos de correspondências bilíngues de dois dicionários fonte

(Inglês, Espanhol)	(Espanhol, Português)
(answer, contestar)	(contestar, contestar) (contestar, responder)

Ao combinarmos estas correspondências por transitividade, construímos um novo dicionário *(Inglês, Português),* com dois novos pares bilingues, um deles errado (marcado mediante asterisco na tabela 2):

Tabela 2: Correspondências bilingues derivadas por transitividade a partir da Tabela 1

(Inglês, Português)
*(answer, contestar)
(answer, responder)

O par errado **(answer, contestar)* foi gerado porque, em Espanhol, *contestar* é um verbo polissémico com dois significados que se lexicalizam em Português de duas maneiras diferentes: *contestar* e *responder*. Finalmente, para podermos identificar o erro, procuramos num corpus comparável Inglês-Português se existe alguma similitude distribucional entre *answer* e *contestar*, por um lado, e *answer* e *responder*, por outro. Este estudo distribucional baseado em corpus permite verificarmos que o par *(answer, responder)* está correto. O dicionário final filtrado não contém o par incorreto: **(answer, contestar)*.

As principais linhas do método apresentado foram descritas em (Gamallo, 2010), onde foi analisado um caso de estudo para a elaboração de um dicionário Inglês-Galego. As principais contribuições do artigo com respeito ao trabalho citado são, por um lado, aplicar o método a um novo par de línguas e, por outro, explicar de modo mais pormenorizado os princípios linguísticos que sustentam o método.

A seguir (seção 2), discutimos alguns trabalhos relacionados, antes de formular, na seção 3, as principais hipóteses linguísticas nas que o nosso método se baseia. Na seção 4, descreve-se em pormenor o nosso método de aprendizagem, para, na seguinte seção (5), apresentar um caso de estudo: a geração de um dicionário bilíngue Inglês-Português. No fim do artigo, comentamos algumas conclusões tiradas do nosso estudo e enumeramos linhas de investigação de trabalhos futuros.

2. Trabalho relacionado

Existem numerosos trabalhos sobre construção automática de novos léxicos bilíngues a partir dos já existentes (Paik et al., 2004; Ahn & Frampton, 2006; Zhang, Ma & Isahara, 2007; Nerima & Wehrli, 2008; Kaji et al., 2008; Simões & Guinovart, 2010; Wehrli et al., 2009).

A maioria destes trabalhos utilizam, para elaborar um novo léxico de correspondências bilíngues, dois dicionários suporte que têm em comum uma língua intermediária. O aspecto crucial desta estratégia é a validação das correspondências aprendidas que são corretas (ou eliminação das espúrias). Em (Nerima & Wehrli, 2008), a validação é feita com ajuda de corpora paralelos, i.e., só são consideradas corretas as correspondências achadas em léxicos alinhados em corpora paralelos.

Outros trabalhos, como o aqui proposto, utilizam extração a partir de corpus não paralelo para realizar a validação (Sammer & Soderland, 2007; Kaji et al., 2008). No entanto, à diferença de trabalhos anteriores sobre extração de léxicos bilíngues a partir de corpora não paralelos (Fung & McKeown, 1997; Fung & Yee, 1998; Rapp, 1999; Chiao & Zweigenbaum, 2002; Shao e Ng, 2004; Saralegui et al., 2008), a nossa proposta utiliza textos comparáveis analisados em dependências. Em (Gamallo, 2008), mostra-se que a precisão dos métodos baseados em dependências para a extração de equivalentes de tradução em corpora comparáveis é maior que a de métodos baseados em *bag-of-words*.

3. Similaridade distribucional *versus* conceptual

O ponto crucial deste trabalho é o processo de validação das correspondências bilíngues derivadas por transitividade, por meio de informação distribucional extraída de corpus. Este processo sustenta-se na seguinte observação: se um par derivado por transitividade também aparece na lista de pares extraídos por similaridade distribucional de corpora comparáveis, então o par é correto. Esta observação é sustentada pelas seguintes conjecturas:

- Nos dicionários bilíngues feitos à mão, cada correspondência bilíngue consta de dois termos que partilham dois aspectos diferentes do significado: os dois têm propriedades *conceptuais* e *distribucionais* similares, é dizer, designam entidades ou conceitos semelhantes (propriedades conceptuais) e combinam-se com entidades ou conceitos semelhantes (distribucionais).

- Nos dicionários bilíngues derivados por transitividade, todas as correspondências derivadas constam de pares de termos que partilham *propriedades conceptuais*, mas nem sempre têm as mesmas *propriedades distribucionais*. Isto é consequência do facto de uma palavra polissêmica ter dois ou mais sentidos relacionados conceptualmente mas não distribucionalmente.

- Nos dicionários extraídos de corpora comparáveis, as correspondências aprendidas constam de pares de termos com as mesmas *propriedades distribucionais*, mas nem sempre partilham *propriedades conceptuais*.

Portanto, se só considerarmos corretos os pares que partilham aspectos conceptuais e distribucionais, então a intersecção dos dicionários derivados por transitividade (semelhança conceptual) com os extraídos de corpora comparáveis (semelhança distribucional) só pode devolver pares bilíngues corretos (semelhança conceptual + distribucional). Este processo de validação é muito preciso. O resultado é um léxico bilíngue limpo, desprovido de erros.

Vejamos um exemplo (figura 1). No dicionário *(Espanhol, Português)*, o nome espanhol *titular* encontra-se nestes dois pares de correspondências: *(titular, manchete)* e *(titular, titular)*. É portanto uma palavra ambígua (polissêmica) associada também a duas traduções no dicionário *(Inglês, Espanhol)*: *(headline, titular)*, *(holder, titular)*. Os dois sentidos de *titular* em Espanhol estão relacionados conceptualmente: nos dois casos, um pequeno rótulo, que pode ser um nome de pessoa ou um título de notícia, serve para designar e identificar um objeto maior, nomeadamente a pessoa titular de uma conta ou a notícia principal de um jornal. Os pares *(Inglês, Português)* derivados por transitividade são: *(headline, manchete)*, **(holder, manchete)*, **(headline, titular)*, *(holder, titular)*, onde a estrelinha marca os pares incorretos. Conjecturamos que os pares são incorretos porque, embora cada um dos dois termos de um par estejam conceptualmente relacionados, não partilham as mesmas propriedades distribucionais. No caso dos pares corretos, os termos de cada par estão relacionados tanto conceptualmente como distribucionalmente.

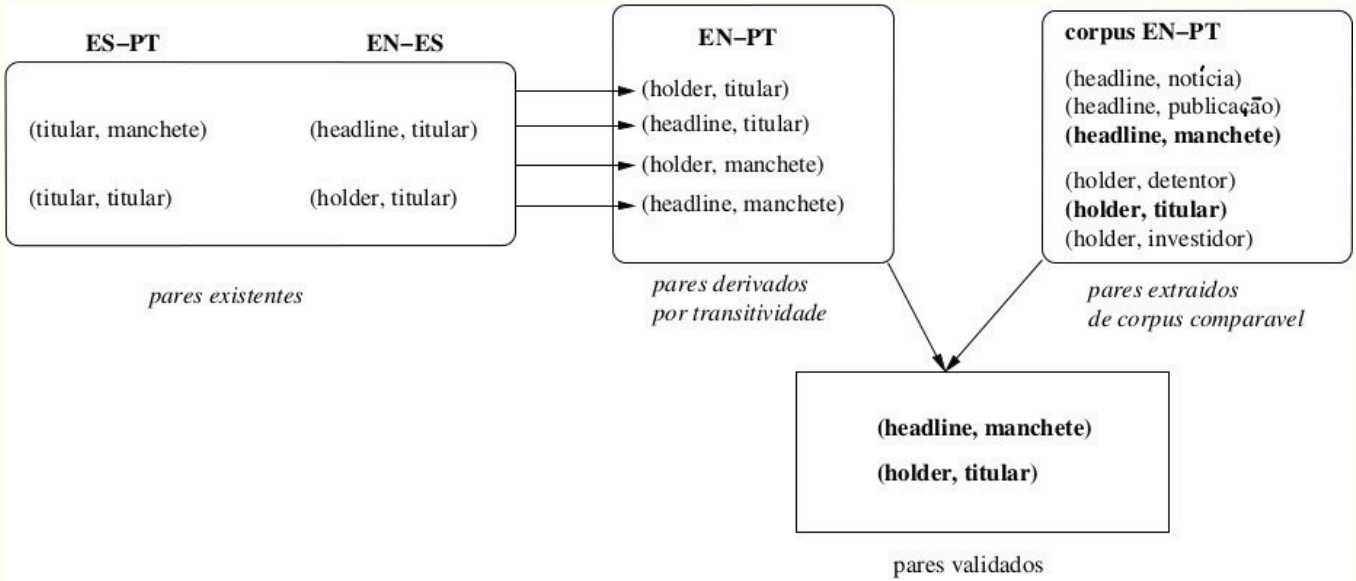

Figura 1: Exemplo de processo de validação

Os pares similares distribucionalmente extraídos de um corpus comparável Inglês-Português permitem identificar, e portanto validar, que pares estão relacionados distribucionalmente. Nas experiências que serão descritas mais a frente, conseguiu-se extrair uma lista de palavras portuguesas similares distribucionalmente a *headline, e que contém* termos como *notícia*, *publicação*, *manchete*, ..., onde todos partilham a mesma distribuição mas só este último descreve o mesmo conceito que *headline*. O termo *titular* não foi extraído, pois a sua distribuição é muito diferente da de *headline*. Analogamente, a lista de termos similares a *holder* está constituída por termos como *detentor*, *titular*, *investidor*, ..., todos com distribuição semelhante (funcionam como agentes), mas só o segundo designa um conceito ou entidade similar a *holder*. Aqui não foi extraído o termo *manchete*, pois a sua distribuição é diferente da de *holder*. A figura 1 mostra de maneira sintética todo o processo de derivação por transitividade e validação através de corpus comparável.

4. Descrição do método

Como já foi dito, a nossa estratégia consiste na execução de duas tarefas: gerar correspondências bilíngues por transitividade e validá-las mediante o uso de pares candidatos extraídos de corpora comparáveis.

4.1. Geração por transitividade

A primeira tarefa foi inspirada pelo trabalho descrito em (Nerima & Wehrli, 2008). Dados dois dicionários bilíngues representados como duas relações (*A*,*B*) e (*B*,*C*), gera-se um dicionário derivado (*A*,*C*) em três etapas:

- Primeiro, cria-se a relação (*A*,*C*)' a partir dos dois dicionários existentes (*A*,*B*) e (*B*,*C*), onde B é a língua pivote ou intermediária. Para cada correspondência bilíngue pertencentes à relação (*A*,*B*), criamos um conjunto de correspondências $\{(a_i, c_1), (a_i, c_2), \ldots, (a_i, c_n)\}$, onde são os termos associados a b_1 em (*B*,*C*). O dicionário derivado final (*A*,*C*)' é o conjunto de todas as novas correspondências geradas.

- Depois, eliminam-se de (A,C)' os pares bilíngues duplicados e obtemos (A,C).
- Finalmente, dividimos (A,C) em dois subconjuntos: $(A,C)_{amb}$, que contém todos os pares com, pelo menos, um elemento derivado de um termo ambíguo de B (língua pivô ou intermediária), e $(A,C)_{unamb}$, contendo só elementos derivados de termos de A não ambíguos. O primeiro subconjunto é uma relação de muitos-para-muitos, enquanto que a segunda é um-para-um.

As tabelas a seguir mostram as três etapas descritas acima. Na tabela 3, mostra-se como se constroem por transitividade os pares *(Inglês, Português)'*, a partir de pares em *(Inglês, Espanhol)* e *(Espanhol, Inglês)*. A tabela 4 mostra a redução dos pares duplicados para obtermos o léxico *(Inglês, Português)*. Finalmente, na tabela 5, separamos os pares com termos ambíguos dos pares sem ambiguidade.

Tabela 3: Primeira etapa: construção por transitividade de *(A, C)'*

(Inglês, Espanhol)	*(Espanhol, Português)*	*(Inglês, Português)'*
(answer, contestar)	(contestar, contestar) (contestar, responder)	(answer, contestar) (answer, responder)
(answer, responder)	(responder, responder)	(answer, responder)
(acquire, adquirir)	(adquirir, adquirir)	(acquire, adquirir)

Tabela 4: Segunda etapa: obtenção de *(A, C)* por redução de *(A, C)'*

(Inglês, Português)'	*(Inglês, Português)*
(answer, contestar) **(answer, responder)** **(answer, responder)** (acquire, adquirir)	(answer, contestar) (answer, responder) (acquire, adquirir)

Tabela 5: Terceira etapa: partição de *(A, C) em* $(A,C)_{amb}$ *e* $(A,C)_{unamb}$

(Inglês, Português)$_{amb}$	*(Inglês, Português)*$_{unamb}$
(answer, contestar) (answer, responder)	(acquire, adquirir)

Como já foi mencionado em (Nerima & Wehrli, 2008), observamos que o dicionário derivado com só palavras não ambíguas, $(A,C)_{unamb}$, é um recurso lexical sem ruido. Na Lexicografia, as palavras com só uma tradução equivalem a termos com pouca ou nula polissemia. Portanto, todas as correspondências derivadas de palavras não ambíguas (um-para-um) são de boa qualidade e podem ser validadas sem ajuda de outros mecanismos. Pelo contrário, como já mostramos mediante o exemplo da palavra espanhola *titular* na seção 3, $(A,C)_{amb}$ é um léxico com ruído.

Na seguinte tarefa, as boas correspondências em $(A,C)_{amb}$ serão identificadas e selecionadas com ajuda de pares similares distribucionalmente extraídos de um corpus comparável com textos em línguas A e C.

4.2. Validação com corpus comparável

O segundo processo é a principal contribuição do nosso trabalho. Consiste em remover as correspondências derivadas de termos ambíguos que não se encontram no léxico de pares similares distribucionalmente, gerado a partir de um corpus comparável anotado com dependências sintáticas. O léxico de pares extraídos do corpus, chamado $(A,C)_{corpus}$, organiza-se desta maneira. Cada termo da língua A, a_i, é associado a uma lista ordenada de termos da língua C, c_1, c_2, ..., c_n, lista que contém os N melhores candidatos de a_i. De maneira inversa, cada termo da língua C, c_i, é associado a uma lista ordenada de termos da língua A, a_1, a_2, ..., a_n, que são os N melhores candidatos de c_1. Portanto, $(A,C)_{corpus}$ é um conjunto de pares bilíngues onde cada par está formado por um termo alvo e um dos seus candidatos de tradução inferido do corpus. Para validar $(A,C)_{amb}$, selecionamos a intersecção entre $(A,C)_{amb}$ e $(A,C)_{corpus}$. A relação resultante é um conjunto de pares bilíngues corretos, mesmo se só contêm termos ambíguos. Finalmente, o dicionário total, chamado $(A,C)_{limpo}$, é a união entre este léxico já validado e o léxico de palavras não ambíguas:

$$(A,C)_{limpo} = (A,C)_{amb} \cap (A,C)_{corpus} \cup (A,C)_{unamb}$$

Na seguinte seção, descrevemos brevemente como se aprende $(A,C)_{corpus}$.

4.3. Extração de equivalentes de tradução a partir de corpora comparáveis

O nosso método para extrair pares similares distribucionalmente a partir de um corpus comparável anotado sintaticamente foi descrito com detalhe em trabalho prévio (Gamallo, 2007). Faremos aqui uma breve introdução ao método. Partimos da seguinte hipótese distribucional:

> *Um lema da língua C é uma tradução candidata do lema da língua A se os contextos léxico-sintáticos onde ocorre são traduções dos contextos léxico-sintáticos onde também ocorre.*

Esta estratégia precisa de uma lista de contextos léxico-sintáticos bilíngues (chamados *contextos semente*) elaborados com ajuda de, por um lado, um dicionário bilíngue existente, (*A*,*C*), e por outro, uma lista de dependências sintáticas comuns nas duas línguas: sujeito, objeto direto, modificação adjetival, complemento preposicional, etc. Com estes elementos, podemos inferir que c_1 é uma tradução candidata de a_1 se os dois termos tendem a ocorrer nos mesmos contextos semente.

Por exemplo, suponhamos que o dicionário (*A*,*C*) contém o par *(subside, baixar)*. Com este par específico de verbos e a dependência *Sujeito*, comum a duas línguas não ergativas como o Inglês e o Português, podemos construir uma correspondência bilíngue entre dois contextos léxico-sintáticos:

<*Sujeito*; *subside*, *NOUN*>
<*Sujeito*; *baixar*, *NOUN*>

Onde <*Sujeito*; *subside*, *NOUN*> é o contexto utilizado para identificar os nomes ingleses que aparecem na posição de sujeito do verbo *subside*, enquanto que <*Sujeito*; *baixar*, *NOUN*> seleciona os nomes portugueses que desempenham o papel de sujeito de *baixar*. As correspondências bilíngues entre pares de contextos léxico-sintáticos assim construídas representam os "contextos semente" nos que se baseia o processo de extração de equivalentes de tradução (ou pares similares distribucionalmente). No nosso exemplo, se os nomes ingleses que aparecem no corpus na posição de sujeito de *subside* são *fever* ou

swelling, os nomes portugueses em posição de sujeito de *baixar* (e.g., *febre* or *inchaço*) são candidatos a serem as suas traduções.

O método de extração consta dos seguintes processos[1]:

- **Parsing multilíngue** Analisam-se os textos nas duas línguas com ajuda do parser multilíngue baseado em dependências, DepPattern[2].
- **Contextos semente** Cria-se uma lista de pares bilíngues de contextos léxico-sintáticos. Utiliza-se, por um lado, o dicionário bilíngue (*A*,*C*) derivado por transitividade e, por outro, um pequeno conjunto de dependências genéricas com o mesmo comportamento nas duas línguas. Devemos realçar aqui que o dicionário bilíngue empregue é o gerado por transitividade contendo todo o ruído derivado dos pares ambíguos sem filtrar.
- ***Hash table*** Uma vez realizada a análise e identificados os contextos semente nos textos, constrói-se uma matriz lema-contexto armazenada em memória como uma *hash table* sem valores nulos (não se tomam em conta os zeros). Cada elemento da hash está constituído por um lema (ou termo multipalavra), um contexto semente e a frequência observada no corpus.
- **Similaridade** Depois, calcula-se o valor de similaridade Dice (Curran & Moens, 2002) entre os pares bilíngues. Para cada lema da língua fonte, selecionamos os *N* mais similares (*N*=10) da língua alvo, os quais representam as suas traduções candidatas.

Na tabela 6, mostramos vários exemplos de pares similares ordenados de maior a menor, extraídos de um corpus comparável Inglês-Português constituído por notícias de jornais. A terceira coluna da tabela mostra o grau de similaridade *Dice* calculado para cada par.

1. Disponível, baixo licença GPL, em: http://gramatica.usc.es/~gamallo/thesaurus/index.htm
2. Disponível, baixo licença GPL, em http://gramatica.usc.es/pln/tools/deppattern.html

Tabela 6: Pares similares distribucionalmente extraídos de corpus comparáveis

Inglês	*Português*	*Coeficiente Dice*
president	presidente	0,61
president	consello	0,49
president	chefe	0,47
president	prefeito	0,46
president	membro	0,46
government	governo	0,63
government	administração	0,54
government	territóri	0,49
government	autoridade	0,49
government	autonomia	0,46
country	país	0,56
country	município	0,51
country	continente	0,39
country	terra	0,37
country	província	0,34

No fim do processo, obtemos a relação $(A,C)_{\text{corpus}}$, que será empregue para validar os pares ambíguos corretos em $(A,C)_{\text{amb}}$. Como foi dito anteriormente, a validação de pares corretos é o resultado da intersecção entre $(A,C)_{\text{corpus}}$ e $(A,C)_{\text{amb}}$.

5. A elaboração de um dicionário Inglês-Português

Para testar a utilidade do nosso método, aplicamo-lo de forma a gerar um novo dicionário ainda inexistente no sistema de tradução automática OpenTrad (Armentano-Oller et al., 2006), nomeadamente o dicionário Inglês-Português. De facto, um dos objetivos a médio-curto prazo das nossas experiências é atualizar os recursos bilíngues de OpenTrad para assim melhorar os resultados do sistema de tradução automática, que é usado por *La Voz de Galicia*, o sexto jornal mais lido na Espanha. Seguiremos um processo similar ao já descrito em (Gamallo, 2010) a respeito da elaboração de um novo dicionário Inglês-Galego integrável em Opentrad. Uma limitação do nosso método é o facto de, por enquanto, só extrair pares de nomes, verbos e adjetivos, pelo que a elaboração do dicionário fica restrita a essas três categorias gramaticais.

5.1. Dicionários existentes e geração por transitividade

O novo dicionário *(Inglês, Português)* é derivado doutros já existentes para os pares de línguas Inglês-Espanhol e Espanhol-Português. Neste experimento, o Espanhol é, portanto, a língua pivô ou intermediária entre o Inglês e o Português. Os dicionários já existentes utilizados como fonte da derivação por transitividade são três:

- **(Inglês_OT, Espanhol_OT)** dicionário Inglês-Espanhol de Opentrad[3], que contém 10.828 pares bilíngues com nomes, adjetivos e verbos, e está disponível com licença livre.
- **(Inglês_CO, Espanhol_CO)** dicionário Inglês-Espanhol de Collins[4], que contém 50.287 pares de nomes, adjetivos e verbos, e só está disponível com licença privativa.
- **(Espanhol_OT, Português_OT)** dicionário Espanhol-Português de Opentrad[5], que contém 10.281 pares de nomes, adjetivos e verbos, disponível com licença livre.

A partir da estratégia descrita na seção 4, foram gerados dois dicionários bilíngues: *(Inglês, Português)_A* e *(Inglês, Português)_B* (ver tabela 7). A primeira linha da tabela mostra os diferentes elementos de *(Inglês, Português)_A*, que foi derivado de dois dicionários de Opentrad: *(Inglês_OT, Espanhol_OT)* e *(Espanhol_OT, Português_OT)*. Contém 5.659 entradas, divididas em dois subconjuntos:

- 1.125 pares ambíguos: *(Inglês, Português)*$\_A_{amb}$
- 4.534 pares não ambíguos: *(Inglês, Português)*$\_A_{unamb}$

Tabela 7: Dicionários derivados por transitividade

3. http://sourceforge.net/projects/apertium/files/apertium-en-es
4. http://www.collinslanguage.com/
5. http://sourceforge.net/projects/apertium/files/apertium-es-pt

dicionários derivados	número de entradas	entradas ambíguas	entradas não ambíguas	dicionários fonte
(Inglês, Português)_A	5.659	1.125	4.534	*(Inglês_OT, Espanhol_OT) (Espanhol_OT, Português_OT)*
(Inglês, Português)_B	10.974	7.310	3.664	*(Inglês_CO, Espanhol_CO) (Espanhol_OT, Português_OT)*
(Inglês, Português)_AB	12.206	7.584	4.622	*(Inglês_CO, Espanhol_CO) (Espanhol_OT, Português_OT) (Espanhol_OT, Português_OT)*

Na segunda linha, o dicionário *(Inglês, Português)_B* é derivado de duas fontes diferentes: um dicionário da Collins, *(Inglês_CO, Espanhol_CO)*, e um outro de Opentrad: *(Espanhol_OT, Português_OT)*. O tamanho do léxico obtido aqui é maior devido ao maior tamanho do dicionário da Collins. Também se observa que o dicionário gerado contém uma proporção muito maior de termos ambíguos, pois já estão presentes no Collins. Nos dicionários de Opentrad a ambiguidade é pequena porque foram construídos para tornar menos complexo o processo de tradução automática.

Finalmente, na terceira linha da tabela, mostra-se o dicionário resultante da união dos dois anteriores: *(Inglês, Português)_AB*. Este vai ser o nosso dicionário de teste. O processo de validação, cujos resultados se descrevem na seguinte seção, estará focado na identificação dos pares ambíguos corretos dentro do subconjunto *(Inglês, Português)_AB*$_{amb}$.

5.2. Corpora comparáveis e validação

Para validar *(Inglês, Português)_AB*$_{amb}$, utilizamos a estratégia de extração de equivalentes de tradução a partir de corpora comparáveis, tal como foi descrito anteriormente nas seções 4.2 e 4.3.

5.2.1. Corpora comparáveis

Construímos dois tipos de corpora. Um primeiro tipo são textos constituídos por notícias de jornais ou de agências de notícias do mesmo período de tempo (ano 2010). O segundo tipo é composto por artigos da Wikipédia sobre o mesmo tópico.

Por um lado, foram compilados mediante *crawling* cinco corpora de notícias monolíngues, três em Inglês e dois em Português, a partir de cinco jornais e agências de notícias: The Guardian, The New York Times, Reuters, Público e Jornal de Notícias. A combinação de todos eles deu lugar a 6 corpora bilíngues comparáveis. A tabela 8 mostra o tamanho (em número de tokens) de cada um dos corpora compilados. Em cada combinação, a tabela mostra o número total de tokens junto com (entre parênteses) o tamanho de cada um dos corpora monolíngues constituintes.

Tabela 8: Tamanho (em milhões de tokens) de seis corpora comparáveis elaborados com cinco fontes de notícias

	Público	**J. de Notícias**
NYT	13.7 (4.6 + 9.1)	16.6 (4.6 + 12)
Guardian	17.3 (8.2 + 9.1)	20.2 (8.2 + 12)
Reuters	21.1 (13 + 9.1)	25 (13 + 12)

Por outro lado, foram também elaborados mais dois corpora comparáveis mediante a extração de artigos da Wikipédia. Mais concretamente, o primeiro corpus foi constituído pelos artigos em Inglês e em Português categorizados, respectivamente, pelos termos "sports" e "desporto". O segundo foi criado selecionando os artigos Ingleses e Portugueses que continham as categorias "country" e "país", respectivamente. Na tabela 9, mostra-se o número de tokens dos dois corpora compilados a partir da Wikipédia.

Tabela 9: Tamanho (em milhões de tokens) dos dois corpora comparáveis elaborados a partir de Wikipédia

	Desporto	**País**
Sport	6.6 (5.1 + 1.6)	
Country		1.2 (0.9 + 0.3)

5.2.2. Extração

Para gerar as listas de equivalentes de tradução, seguimos o método de extração descrito na seção 4.3. Primeiro, os textos foram analisados

sintaticamente com DepPattern de forma a extrair as dependências entre lemas. Só tomamos em conta as dependências que contêm verbos, nomes, ou adjetivos. A entrada de DepPattern é texto anotado com o etiquetador morfossintático FreeLing (Carreras et al. 2004). A seguir, com ajuda do dicionário derivado por transitividade, *(Inglês, Português)_AB*, e de regras sintáticas de correspondências bilíngues, geramos a lista semente de contextos léxico-sintáticos bilíngues. Como nesta tarefa, de natureza estatística, é preferível dar mais importância à cobertura do que à precisão, utilizamos o dicionário com pares ambíguos e, portanto, não deitamos fora os pares errados. Posteriormente, com base nas dependências e na lista semente de contextos bilíngues, construímos 8 matrizes (uma por corpus) onde as dimensões são os contextos bilíngues e os objetos os lemas nas duas línguas. Finalmente, calculamos um valor de similaridade entre todos os pares de palavras dentro de cada uma das matrizes. Para cada lema inglês, selecionamos os 10 portugueses com um grau maior de similaridade, considerados como possíveis traduções do inglês. Como a similaridade é uma relação assimétrica, fizemos o mesmo desde o Português para o Inglês. No final do processo, obtivemos 8 dicionários estatísticos de equivalentes de tradução:

(Inglês, Português)$_{nyt\text{-}publ}$ dicionário extraído a partir do New York Times e Público
(Inglês, Português)$_{nyt\text{-}jn}$ dicionário extraído a partir do New York Times e Jornal de Notícias
(Inglês, Português)$_{guar\text{-}publ}$ dicionário extraído a partir de The Guardian e Público
(Inglês, Português)$_{guar\text{-}jn}$ dicionário extraído a partir de The Guardian e Jornal de Notícias
(Inglês, Português)$_{reut\text{-}publ}$ dicionário extraído a partir de Reuters e Público
(Inglês, Português)$_{reut\text{-}jn}$ dicionário extraído a partir de Reuters e Jornal de Notícias
(Inglês, Português)$_{sport\text{-}desp}$ dicionário extraído a partir de artigos de desporto da Wikipédia inglesa e portuguesa
(Inglês, Português)$_{country\text{-}país}$ dicionário extraído a partir de artigos sobre países da Wikipédia inglesa e portuguesa

5.2.3. Validação

De forma a validar a correção dos pares derivados por transitividade e formados por lemas ambíguos, realizamos a simples intersecção entre

esse conjunto de pares ambíguos e os dicionários estatísticos derivados de corpora comparáveis. A tabela 10 mostra o resultado de intersectar cada um dos 8 dicionários estatísticos com o conjunto *(Inglês, Português)_AB*$_{amb}$, que contém 8.620 entradas. A primeira linha da primeira coluna da tabela mostra o valor obtido pela intersecção do léxico ambíguo *(Inglês, Português)_AB*$_{amb}$ com o *(Ingl,Port)*$_{nyt\text{-}publ}$: 1.147 lemas, que representam o 15% do total de pares ambíguos. Nas linhas restantes, aparecem os valores da intersecção com os outros dicionários. Além do número absoluto de palavras intersectadas (quer dizer validadas), a tabela mostra também a percentagem de palavras validadas (entre parênteses) em relação com o número total de entradas no léxico original de lemas ambíguos.

Tabela 10: Validação do léxico ambíguo mediante 8 corpora comparáveis

	(Ingl, Port)_AB$_{amb}$	**tamanho acumulado**
(Ingl, Port)$_{nyt\text{-}publ}$	1147 (15%)	1147 (15%)
(Ingl, Port)$_{nyt\text{-}jn}$	1069 (14%)	1384 (18%)
(Ingl, Port)$_{guar\text{-}publ}$	1488 (20%)	1850 (24%)
(Ingl, Port)$_{guar\text{-}jn}$	1382 (18%)	2030 (27%)
(Ingl, Port)$_{reut\text{-}publ}$	1384 (18%)	2183 (29%)
(Ingl, Port)$_{reut\text{-}jn}$	1264 (17%)	2247 (30%)
(Ingl, Port)$_{sport\text{-}desp}$	695 (9%)	2367 (31%)
(Ingl, Port)$_{count\text{-}país}$	434 (6%)	**2411 (32%)**

A segunda coluna representa os valores acumulados das intersecções, sendo a última linha o valor final atingido (2411 - 32%) pela união das 8 intersecções. Como se pode observar, a união sucessiva das intersecções vai aumentando o número de validações até ao 32% do total. Este valor é superior ao obtido (26%) em (Nerima & Wehrli, 2008), onde se utilizaram corpora paralelos para validar as correspondências ambíguas de um dicionário Inglês-Alemão derivado por transitividade. Além de obter melhores resultados percentuais, o facto de nós usarmos corpora comparáveis, mais fáceis de encontrar na Web do que os paralelos, deixa aberta a possibilidade de podermos atingir, com facilidade, valores ainda superiores com a exploração de novos corpora.

Por último, devemos sublinhar que a qualidade das correspondências validadas é muito alto. O número de erros é muito pequeno e próximo ao achado em dicionários elaborados manualmente. Isto prova que as hipóteses formuladas na seção 3 parecem corretas.

5.3. O léxico final sem ruído

No fim do processo, fazemos a união dos pares validados com o dicionário de pares não ambíguos (i.e., as correspondências uma-para-uma): *(Ingl, Port)_AB*$_{unamb}$. A tabela 11 mostra o número de entradas obtido em cada uma das etapas do processo. A primeira linha representa o dicionário inicial com ruido, com 12.206 entradas, gerado por transitividade antes do filtrado. A segunda linha representa o total de pares validados com ajuda dos corpora comparáveis: 2.411. A terceira linha mostra o número de pares não ambíguos: 4.622. E a última linha representa o número total de pares limpos, 7.033, que o nosso método conseguiu gerar. Este número representa o 58% do total de pares com ruido, 12.206, resultado da derivação por transitividade.

Tabela 11: Processo de construção do *dicionário limpo final* Inglês-Português

(Inglês,Português)	*número de entradas*
OpenTrad + Collins	12.206
pares validados	2.411
pares não ambíguos	4.622
dicionário limpo total	7.033 (58%)

Em resumo, o dicionário final é o resultado das seguintes operações de conjuntos:

$$
\begin{aligned}
&\textbf{\textit{(Ingl, Port)\_AB}}_{\textbf{\textit{limpo}}} = \\
&(\ \textit{(Ingl, Port)\_A}_{amb} \cup \textit{(Ingl, Port)\_B}_{amb}\ ) \cap \\
&(\textit{(Ingl, Port)}_{nyt\text{-}publ} \cup \textit{(Ingl, Port)}_{nyt\text{-}jn} \cup \\
&\textit{(Ingl, Port)}_{guar\text{-}publ} \cup \textit{(Ingl, Port)}_{guar\text{-}jn} \cup \\
&\textit{(Ingl, Port)}_{reut\text{-}publ} \cup \textit{(Ingl, Port)}_{reut\text{-}jn} \cup \\
&\textit{(Ingl, Port)}_{sport\text{-}desp} \cup \textit{(Ingl, Port)}_{countr\text{-}pais}\ ) \cup \\
&(\ \textit{(Ingl, Port)\_A}_{unamb} \cup \textit{(Ingl, Port)\_B}_{unamb}\ )
\end{aligned}
$$

É importante sublinhar que o dicionário final, mesmo se só contém 58% das entradas geradas por transitividade, o seu tamanho não fica muito longe do tamanho do mais pequeno dos dicionários fonte, *(Espanhol_OT, Português_OT)*, que contém 10.281 entradas.

Finalmente, a tabela 12 mostra um excerto de mais de 50 entradas do dicionário limpo. A terceira coluna representa a categoria

morfossintáctica de cada par: NOUN (nome), VERB (verbo) e ADJ (adjectivo).

Tabela 12: Breve excerto do dicionário limpo final

Inglês	Português	Categoria
stroke	acariciar	VERB
stroll	passeio	NOUN
stroll	passear	VERB
strong	forte	ADJ
structural	estrutural	ADJ
structure	estrutura	NOUN
struggle	contenda	NOUN
struggle	luta	NOUN
struggle	brigar	VERB
struggle	lutar	VERB
student	estudante	NOUN
student	aluno	NOUN
studio	estúdio	NOUN
study	estúdio	NOUN
study	estudo	NOUN
study	estudar	VERB
study_thoroughly	afundar	VERB
stuff	rechear	VERB
stupid	idiota	ADJ
stupid	estúpido	ADJ
stupidity	estupidez	NOUN
stupidity	burrice	NOUN
style	estilo	NOUN
sub-Saharan	subsaariano	ADJ
subhead	subtítulo	NOUN
subject	sujeito	ADJ
subject	assunto	NOUN
subject	matéria	NOUN
subject	sujeito	NOUN
subject	tema	NOUN
subject	submeter	VERB
sublime	sublime	ADJ
submerge	afundar	VERB
submerge	submergir	VERB
submission	entrega	NOUN
submit	submeter	VERB
submit	apresentar	VERB
submit	entregar	VERB
subpoena	citar	VERB
subscription	assinatura	NOUN
subsequent	posterior	ADJ
subsidy	subsídio	NOUN
substantial	substancial	ADJ
substantial	importante	ADJ
substantial	considerável	ADJ
substantive	fundamental	ADJ
substitute	substituir	VERB
subtitle	subtítulo	NOUN
subtle	subtil	ADJ
subtlety	subtileza	NOUN
subtlety	sutileza	NOUN
substract	restar	VERB
subway	metro	NOUN
success	sucesso	NOUN
successive	sucessivo	ADJ
successor	sucessor	NOUN
sudden	brusco	ADJ
sudden	súbito	ADJ
sudden	repentino	ADJ

6. Conclusões e Trabalho Futuro

O método lexicográfico proposto neste artigo é totalmente automático. O novo léxico gerado não precisa de nenhum tipo de revisão manual pois a qualidade dos pares validados é muito alta, semelhante à atingida por lexicógrafos humanos. A principal contribuição do método é a utilização, para a validação das correspondências derivadas por transitividade, de equivalentes de tradução extraídos de corpora comparáveis anotados sintaticamente. O caso de estudo apresentado mostrou que para alargar a cobertura do dicionário, sem perder precisão, só é preciso ir à procura de mais fontes de informação que nos forneçam mais textos comparáveis.

O principal problema do método é o facto de ser dependente da língua, pois requer de um parser sintático para anotar o corpus. No entanto, por forma a tratar o maior número de línguas possível, utilizamos um analisador multilíngue robusto, DepPattern, facilmente adaptável a mais línguas.

Em trabalho futuro, elaboraremos um sistema de atualização automática dos dicionários do sistema de tradução Opentrad, que permita converter e adaptar os léxicos gerados com o nosso método para o formato requerido por Opentrad.

Recebido em novembro de 2011
Aprovado em outubro de 2013
E-mail: pablo.gamallo@usc.es

Referências bibliográficas

Ahn, Kisuh & Matthew Frampton. 2006. Automatic generation of translation dictionaries using intermediary languages. In *Cross-Language Knwoledge Induction Workshop of the EACL'06*, p. 41-44, Trento, Italy.

Armentano-Oller, Carme; Rafael C. Carrasco; Antonio M. Corbí-Bellot; Mikel L. Forcada; Mireia Ginestí-Rosell; Sergio Ortiz-Rojas; Juan Antonio Pérez-Ortiz; Gema Ramírez-Sánchez; Felipe Sánchez-Martínez & Miriam A. Scalco. 2006. Open-source Portuguese-Spanish machine translation. In *Lecture Notes in Computer Science, 3960*, p. 50–59.

Carreras, X.; I. Chao; L. Padró & M. Padró. 2004. An Open-Source Suite of Language Analyzers. In *4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.

Chiao, Y-C. & P. Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *19th COLING'02*.

Curran, James R. & Marc Moens. 2002. Improvements in Automatic Thesaurus Extraction. In *ACL Workshop on Unsupervised Lexical Acquisition*, p. 59–66, Philadelphia.

Fung, Pascale & Kathleen McKeown. 1997. Finding terminology translation from non-parallel corpora. In *5th Annual Workshop on Very Large Corpora*, p. 192–202, Hong Kong.

Fung, Pascale & Lo Yuen Yee. 1998. An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In *Coling'98*, p. 414–420, Montreal, Canada.

Gamallo P. 2007. Learning Bilingual Lexicons from Comparable English and Spanish Corpora. In *Proceedings of Machine Translation Summit XI* , Copenhagen, Denmark, pp. 191-198.

_____. 2008. Evaluating two different methods for the task of extracting bilingual lexicons from comparable corpora, In *LREC 2008 Workshop on Comparable Corpora*, Marrakech, Marroco, pp. 19-26.

_____. 2010. Automatic Generation of Bilingual Dictionaries Using Intermediary Languages and Comparable Corpora , *Lecture Notes in Computer Science*, vol. 6008, Springer-Verlag, 473-483.

Kaji, Hiroyuki; Shin'ichi Tamamura, & Dashtseren Erdenebat. 2008. Automatic construction of a japanese-chinese dictionary via english. In *LREC'08*, Marrakesh, Marocco.

Nerima, Luka & Eric Wehrli. 2008. Generating bilingual dictionaries by transitivity. In *LREC'08*, p. 2584–2587, Marrakesh, Marocco.

Kyonghee Paik; Satoshi Shirai & Hiromi Nakaiwa. 2004. Automatic construction of a transfer dictionary considering directionality. In *20th International Conference on Computational Linguistics*, p. 31–38, Geneva, Switzerland.

Rapp, Reinhard. 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. In *ACL'99*, p. 519–526.

Sammer, M. & S. Soderland. 2007. Building a sense-distinguished multilingual lexicon from monolingual corpora and bilingual lexicons. In *Machine Translation Summit XI*.

Saralegui, X.; I. San Vicente & A. Gurrutxaga. 2008. Automatic generation of bilingual lexicons from comparable corpora in a popular science domain. In *LREC 2008 Workshop on Building and Using Comparable Corpora*.

Shao, Li & Hwee Tou Ng. 2004. Mining New Word Translations from Comparable Corpora. In *20th International Conference on Computational Linguistics (COLING 2004)*, p. 618–624, Geneva, Switzerland.

Simões, Alberto & Xabier Guinovart. 2010. Translation Dictionaries by Triangulation. Fala 2010 II Iberian SLTech Workshop, p. 171-174.

Wehrli, Eric; Luka Nerima & Yves Scherrer. 2009. Deep linguistic multilingual translation and bilingual dictionaries. In *4th Workshop on Statistical Machine Translation*, p. 90–94, Athens, Greece.

Zhang, Yujie; Quing Ma & Hitoshi Isahara. 2007. Building japanese-chinese translation dictionary based on EDR japanese-english bilingual dictionary. In *Machine Translation Summit XI*, p. 699–706.