Universidade Nova de Lisboa
Faculdade de Ciências e Tecnologia
Departamento de Informática

Dissertação de Mestrado

Mestrado em Engenharia Informática

Suporte à extracção de traduções de termos simples ou compostos em ambiente multilingue

Nº 30095 Gonçalo Nuno Ramos Gomes

2º Semestre de 2009/2010

28 de Julho de 2010

Universidade Nova de Lisboa
Faculdade de Ciências e Tecnologia
Departamento de Informática

Dissertação de Mestrado

Suporte à extracção de traduções de termos simples ou compostos em ambiente multilingue

Nº 30095 Gonçalo Nuno Ramos Gomes

Orientador: Prof. Doutor José Gabriel Pereira Lopes

Trabalho apresentado no âmbito do Mestrado em Engenharia Informática, como requisito parcial para obtenção do grau de Mestre em Engenharia Informática.

2º Semestre de 2009/2010

28 de Julho de 2010

Agradecimentos

Quero agradecer em primeiro lugar ao meu orientador, o Professor Doutor Gabriel Pereira Lopes, pela disponibilidade, pelo incentivo e pelas inúmeras vezes me ajudou a encontrar o rumo certo nas minhas investigações.

Agradeço aos meus pais o muito amor que me deram, os valores que me transmitiram e o facto de me incentivarem a dar o melhor de mim.

Agradeço também às minhas irmãs o amor a amizade e o companheirismo que sempre tivemos, bem como o apoio "logístico" dado durante a escrita desta dissertação.

Agradeço muito às minhas avós que sempre se orgulharam de mim, e das quais tenho saudades por já não se encontrarem entre nós. Sempre estiveram presentes em todos os momentos importantes da minha vida, pelo que esta é a primeira vez que não tenho pelo menos uma das avós a dar-me o seu apoio incondicional.

Resumo

A construção automática de léxicos bilingues é indispensável para aplicações como o acesso a informação disponível em várias línguas, a tradução automática, a construção de ontologias multilingues, entre outras. Nos últimos anos tem havido alguma actividade no sentido de extrair traduções auxiliadas por léxicos bilingues existentes entre línguas com maiores recursos, X-Y e Y-Z, por exemplo, para pares de línguas com menores recursos, X-Z, utilizando uma das línguas com mais recursos (a língua Y, no caso) como língua pivô.

Neste trabalho, assume-se o Português (PT) como língua pivô. Dada a existência de um léxico de traduções de palavras e de multi-palavras minimamente desenvolvido, com cerca de 200.000 entradas, entre Português e Inglês (EN), partindo da semelhança estrutural e lexical das línguas Portuguesa e Espanhola (ES), extraiu-se traduções de palavras para o par EN-ES, utilizando um corpus de textos paralelos (dois textos são paralelos se forem tradução um do outro ou ambos forem tradução de um mesmo texto fonte), existente para todas as línguas da União Europeia, detectando primeiro as palavras que possam ser cognatas (semelhantes na forma e com o mesmo significado). Considerando as entradas do léxico bilingue PT-EN, identificou-se as possíveis traduções em Inglês das palavras espanholas cognatas com palavras portuguesas previamente identificadas. Em seguida, os possíveis cognatos entre Português e Espanhol e as possíveis traduções de Espanhol em Inglês foram avaliadas adicionalmente quanto à sua semelhança nos textos da colecção em que ocorrem, recorrendo a medidas de semelhança utilizadas para estes efeitos e definindo um limiar de semelhança para a aceitação dos cognatos entre PT e ES como traduções e das traduções inferidas, entre o Espanhol e o Inglês, via o léxico bilingue PT-EN. Os resultados foram avaliados manualmente.

Conseguiu-se aumentar a produtividade dos avaliadores das traduções de termos, passando-lhes traduções com um elevado grau de precisão. Discute-se, os resultados obtidos fazendo variar: o grau de semelhança entre possíveis cognatos, a medida de semelhança entre termos de línguas diferentes, distinguindo-se também o tipo de contextos em que ocorrem (textos completos paralelos, frases paralelas ou segmentos mais curtos obtidos pelo alinhamento realizado) e os limiares de aceitação utilizados.

Palavras-Chave: tradução automática, alinhamento de textos paralelos, extracção de equivalentes de tradução, geração automática de léxicos multilingues, língua pivô

Abstract

The automatic construction of bilingual lexicons is essential for applications such as access to information available in several languages, machine translation, and construction of multilingual ontologies, among others. In recent years there has been some activity in order to extract translations using bilingual lexicons between languages with greater resources, X-Y and Y-Z, to language pairs with fewer resources, X-Z, using a language with more resources (the language Y in this case) as their pivot.

In this work, we assume Portuguese (PT) as a bridge language. Given the existence of a translation lexicon, with about 200,000 Portuguese and English (EN) entries, based on Portuguese and Spanish (ES) lexical and structural similarity, translations of words for the EN-ES pair are extracted, using a corpus of parallel texts (two texts are parallel if one is translation of the other or both are translating the same source text), existing for all EU languages, detecting first the probable cognate words (similar in form and with the same meaning). Considering the bilingual lexicon PT-EN entries, are identified the possible English translations of Spanish words previously identified as Portuguese words cognates. Then the possible cognates between Spanish and Portuguese and the possible translations of Spanish into English will be evaluated further regarding to their similarity in the collection of texts in which they occur, using a similarity measure for these purposes and setting a threshold for similarity acceptance of PT-ES cognates as translations and for the inferred translations between Spanish and English, via the bilingual lexicon PT-EN. The results were evaluated manually.

We were able to increase the human evaluator's productivity by passing them translations with a high degree of accuracy. The obtained results were discuss by varying: the similarity threshold between possible cognates, the measure of similarity between terms of different languages, distinguishing also the type of contexts in which they occur (parallel complete texts, parallel phrases or shorter segments obtained by the alignment performed) and the acceptance thresholds used.

Keywords: machine translation, parallel texts alignment, extraction of translation equivalents, multilingual lexical database generation, pivot language

Índice

Lista de Figuras

Lista de Tabelas

1. Introdução

A construção automática de léxicos bilingues é indispensável para aplicações várias de que destaco o acesso a informação disponível em várias línguas, a tradução automática, a construção de ontologias multilingues, entre outras. Nos últimos anos tem havido alguma actividade no sentido de extrair traduções auxiliadas por léxicos bilingues existentes entre línguas com maiores recursos, X-Y e Y-Z, por exemplo, para pares de línguas com menores recursos, X-Z, utilizando uma das línguas com mais recursos (a língua Y, no caso) como língua pivô [Gideon, Mann et al., 2001; Shafer, Charles et al., 2002].

Neste trabalho, foi assumido o Português (PT) como língua pivô, dada a existência de um léxico de traduções de palavras simples e de multi-palavras minimamente desenvolvido, com cerca de 200.000 entradas, entre Português (PT) e Inglês (EN). Partindo da semelhança estrutural e lexical das línguas Portuguesa e Espanhola (ES), foram extraídas traduções de palavras para o par EN-ES, utilizando um corpus de textos paralelos[1], existente para todas as línguas da União Europeia (corpus da Constituição Europeia usado na tese de Mestrado de Luís Gomes [Gomes, Luís, 2009]), detectando primeiro as palavras que possam ser cognatas (semelhantes na forma e com o mesmo significado). Considerando as entradas do léxico bilingue PT-EN, foram identificadas as possíveis traduções em Inglês das palavras espanholas extraídas como cognatas das palavras portuguesas previamente identificadas na colecção de textos paralelos utilizada. Em seguida, os possíveis cognatos entre Português e Espanhol e as possíveis traduções de Espanhol em Inglês foram avaliadas suplementarmente nos textos da colecção em que ocorrem. Para isso recorreu-se a medidas de semelhança utilizadas para estes efeitos [ver secção 2.3] e foi definido um limiar de semelhança para a aceitação dos cognatos entre PT e ES como traduções e das traduções inferidas, entre o Espanhol e o Inglês, via o léxico bilingue PT-EN.

Tendo em linha de conta a capacidade instalada para alinhar textos paralelos [Ribeiro, António, 2002; Gomes, Luís, 2009][2], utilizando os léxicos PT-ES e EN-ES entretanto

[1] Dois textos são paralelos se forem tradução um do outro ou se ambos forem tradução de um mesmo texto fonte.

[2] O processo de alinhamento, na perspectiva destes autores, ao estabelecer correspondências muito seguras entre traduções de termos, mono ou multi-palavra, divide cada um dos textos paralelos em segmentos que, desejavelmente, deveriam continuar a ser paralelos, i.e. traduções uns dos outros. Observa-se porém que a precisão dos alinhamentos obtidos ao nível destes segmentos, estando longe de ser de 100%, aumenta no entanto

extraídos, alinharam-se os corpus paralelos PT-ES e EN-ES, fazendo em seguida a extracção de novas traduções utilizando o extractor descrito em [Lopes, Gabriel e Aires, José, 2009]. Os resultados obtidos em qualquer das fases foram avaliados manualmente. Com este trabalho, foi possível aumentar a produtividade dos avaliadores das traduções extraídas, passando-lhes para validação traduções com um elevadíssimo grau de precisão.

1.1Descrição e contexto

O alinhamento de textos paralelos e a criação de léxicos bilingues (constituídos pelas traduções de termos simples ou compostos entre duas línguas) [Lopes, Gabriel e Aires, José, 2009] que permitem melhorar interactivamente esses mesmos alinhamentos [Gomes, Luís, 2009] é um processo que requer algum processamento e muito esforço humano na validação das respectivas entradas, esforço este que importa diminuir tanto quanto possível.

Com efeito, a utilização de léxicos obtidos automaticamente pressupõe que os mesmos sejam validados, permitindo a partir daí um incremento qualitativo dos alinhamentos e, consequentemente, das traduções obtidas à posteriori a partir desses novos alinhamentos, mais precisos e mais finos. Para esta tarefa, é necessário recorrer a equipas de linguistas que efectuem a validação, e correcção das inúmeras entradas extraídas automaticamente, recorrendo sempre que necessário a *concordancers*, para verificarem se cada uma das traduções extraídas é de facto aceitável.

Desta forma, partindo de um léxico bilingue entre duas línguas X e Y e dos textos paralelos de dois pares de línguas X-Z e Y-Z, foi desenvolvido um processo de extracção de traduções para os pares de línguas X-Z e Y-Z, que seja conduzido da forma mais eficiente. Conseguiu-se um ganho substancial no tempo de processamento e no esforço humano dispendido nas validações das entradas para estes dois novos pares de línguas.

Foram escolhidos os pares Português-Espanhol (PT-ES) e Inglês-Espanhol (EN-ES), utilizando o léxico bilingue Português-Inglês (PT-EN), entretanto extraído e validado, como alavanca para melhorar a qualidade e velocidade de avaliação dos léxicos PT-ES e EN-ES extraídos automaticamente.

Actualmente não dispomos de nenhum destes léxicos PT-ES e EN-ES e como consequência disso, os alinhamentos dos textos destes dois pares de línguas, não têm a qualidade dos alinhamentos correspondentes do par EN-PT.

---

com a quantidade e a qualidade do conhecimento disponível sobre a tradução de palavras e multi-palavras. Como é relatado em [Gomes, Luís et al., 2009], a precisão destes alinhamentos subfrásicos passou de um máximo de 75,5% (quando se utilizaram como alinhadores apenas possíveis cognatos entre PT e EN [Bilbao, Darriba et al., 2005]) para 84,5% quando se utilizou um léxico bilingue então com cerca de 60.000 entradas automaticamente extraídas, validadas manualmente, para o mesmo par de línguas.

1.2Solução apresentada

O *Acquis Communautaire* é uma colecção de textos paralelos contendo as leis aplicáveis a cada estado membro. De acordo com a política multilingue da Comissão Europeia, estes textos são publicados nas seguintes 23 línguas de estados membros: búlgaro, checo, dinamarquês, alemão, grego, inglês, irlandês, espanhol, estónio, finlandês, francês, húngaro, italiano, lituano, letão, maltês, neerlandês, polaco, português, romeno, eslovaco, esloveno e sueco. Este é dos maiores corpus paralelos existentes, quer pelo tamanho quer pela quantidade de línguas que o constituem. Devido ao seu tamanho, para realizar o trabalho de investigação proposto neste documento, utilizei antes o corpus da Constituição Europeia (usado na tese de Mestrado de Luís Gomes [Gomes, Luís, 2009]), escrito também nas mesmas línguas, por ser mais pequeno, facilitando desse modo a validação dos resultados obtidos.

O léxico EN-PT é neste momento o mais desenvolvido e utilizado, pois serve o propósito de vários trabalhos em curso no DI/FCT/UNL. Consequentemente, tem vindo a ser enriquecido e validado, pelo que, no início desta tese existia um léxico com cerca de 180.000 entradas constituídas por traduções de termos simples e compostos, ultrapassando neste momento as 250.000 entradas.

Dada a similaridade existente entre as línguas Portuguesa e Espanhola, quer no plano vocabular quer no plano estrutural e gramatical, que permite que cerca de 35% das palavras tenha exactamente a mesma grafia (são palavras homógrafas) [Ribeiro, António et al., 2000], enquanto o número de homógrafas entre Português e Inglês se fica pelos 15%, de acordo com a mesma publicação. Estes factos que motivaram a escolha do par de línguas Português e Espanhol para fazer a identificação de possíveis cognatos[3] na colecção de textos paralelos seleccionada (Constituição Europeia), de forma a obter um léxico inicial de traduções com um elevado grau de precisão. Para esta tarefa foi usada a distância de *Levenshtein* [ver secção 2.3.1], que permite obter directamente um grau de dissemelhança ou correspondência entre os termos portugueses e espanhóis, já que na sua grande maioria estes diferem apenas em algumas letras.

A distância de *Levenshtein* é uma métrica que avalia a diferença entre duas cadeias de caracteres, medindo-a através do número mínimo de operações de edição[4] necessárias para transformar uma cadeia na outra. Se queremos identificar as formas de palavras em línguas diferentes que sejam o mais semelhantes possível, temos de limitar o número destas operações de edição até um determinado limiar como seja admitir, por exemplo, no máximo uma operação de edição. Se fizermos isto sem outras precauções, como seja a

[3] Palavras semelhantes na forma e com o mesmo significado.

[4] Operações de inserção, de remoção, ou de substituição de um único carácter

de ter em linha de conta o tamanho das palavras que estamos a comparar, podemos chegar a resultados de baixa qualidade para os objectivos que pretendemos atingir. Nos dois textos paralelos abaixo (Figura 1), depois de realizadas algumas operações de normalização, como seja reduzir todos os caracteres ao mesmo tipo (todos em minúsculas, por exemplo) e separar as contracções das preposições com artigos portugueses nos seus elementos constitutivos (preposições e artigos correspondentes), como acontece ao substituir "no" por "em o", "à" por "a a", etc., poderia assumir-se que:

- a palavra portuguesa "a" é semelhante às palavras espanholas "la", "a" e "ha";
- a palavra portuguesa "um" é semelhante à palavra espanhola "un"
- a palavra portuguesa "ibero-americana" é semelhante à palavra espanhola "iberoamericana"
- a palavra portuguesa "substancial" é semelhante à palavra espanhola "sustancial"
- a palavra portuguesa "ano" é semelhante à palavra espanhola "año"
- a palavra portuguesa "em" é semelhante à palavra espanhola "en"
- a palavra portuguesa "âmbito" é semelhante à palavra espanhola "ámbito"
- a palavra portuguesa "meio" é semelhante à palavra espanhola "medio"
- as palavras portuguesas "de", "madrid", "anterior", "guadalajara","México", "se", "programas", "concretos" "especialmente", "cultura", "mundo","indígena" "ambiente" são homógrafas das correspondentes palavras espanholas
- Mas as palavras "educação" e "saúde" diferem das palavras "educación" e "salud" por um número de operações de edição superiores a uma.

La Cumbre Iberoamericana de Madrid ha significado un sustancial avance en relación a la celebrada el año anterior en Guadalajara, (México), pues se han aprobado programas concretos, especialmente en el ámbito de la educación y cultura, salud, mundo indígena y medio ambiente.	A Cimeira Ibero-Americana de Madrid constituiu um progresso substancial relativamente a a cimeira realizada em o ano anterior em Guadalajara, em o México, porquanto se aprovaram programas concretos, especialmente em o âmbito de a educação e de a cultura, de a saúde, de o mundo indígena e de o meio ambiente.

Figura 1: Exemplo de texto paralelo ES-PT

Usando o texto da Figura 1 como exemplo, de entre os anteriores possíveis cognatos se eliminarmos numa primeira avaliação aqueles que tenham menos de três caracteres, ou seja palavras como “a”, “um”, “em”, “de” e “se”, o alinhamento dos dois textos que se obtém, assumindo como correctas as cognatas antes extraídas, é o que se apresenta na Tabela 1, onde cada um dos segmentos obtidos está numerado à esquerda e, na coluna do centro, os dois asteriscos assinalam que os dois lados do alinhamento foram assumidos como traduções.

1	A cimeira		La cumber
2	Ibero-americana	**	Iberoamericana
3	De		de
4	Madrid	**	Madrid
5	constituiu um progresso		ha significado un
6	Substancial	**	sustancial
7	relativamente a a cimeira realizada em o		avance en relación a la celebrada el
8	Ano	**	año
9	Anterior	**	anterior
10	Em		en
11	Guadalajara	**	Guadalajara
12	,	**	,
13	em o		(
14	México	**	México
15			)
16	,		,
17	porquanto se aprovaram		pues se han aprobado
18	Programas	**	programas
19	Concretos	**	concretos
20	,	**	,
21	Especialmente	**	especialmente
22	em o		en el
23	Âmbito	**	ámbito
24	de a educação e de a		de la educación y
25	Cultura	**	cultura
26	,	**	,
27	de a saúde		salud
28	,	**	,
29	de o		
30	Mundo	**	mundo
31	Indígena	**	indígena

32	e de o		y
33	Meio	**	medio
34	Ambiente	**	ambiente
35	.	**	.

Tabela 1: Alinhamento dos dois textos usando as palavras cognatas

Tomando o alinhamento PT-ES acima apresentado, pode extrair-se com alguma segurança mais traduções além das que já foram utilizadas para alinhar aqueles textos. Assim sendo, para este processo, e apenas para ilustrar a metodologia a utilizar, recorre-se à medida de *Dice* reproduzida a seguir [ver também a secção 2.3.4].

$$Dice(X,Y) = 2 \cdot f(X,Y) / (f(X) + f(Y))$$

Esta medida, dá-nos a semelhança entre duas unidades textuais X e Y através da frequência da coocorrência de X e Y nos mesmos segmentos alinhados (denotada por $f(X,Y)$) e da frequência separada de X e de Y nos textos das línguas respectivas.
Por exemplo, a possibilidade de "em" ser tradução de "en" é medida por $dice("em","en") = 2 \times 3 / (4+3) = 0.857$. A possibilidade de "a" ser tradução das palavras espanholas "la", "a" e "ha" é medida pelos 3 coeficientes de *Dice* seguintes:

$dice("a","la") = 2 \times 3 / (6+3) = 0.666$;
$dice("a","a") = 2 \times 1 / (1+6) = 0.286$;
$dice("a","ha") = 2 \times 0 / (1+6) = 0$;

A possibilidade de "um" ser tradução de "un" é medida pelo seguinte valor do coeficiente de *Dice*: $dice("um","un") = 2 \times 1 / (1+1) = 1$. Conseguir-se-ia também, calcular os valores $dice("se","se") = 1$, $dice("educação","educación") = 1$ e $dice("saúde","salud") = 1$. Considerando um limiar fixo agora arbitrariamente[5] em 0.6, poder-se-ia extrair adicionalmente as traduções correctas: "cimeira←→cumbre", "educação←→educación", "saúde←→salud", "a←→la", "e←→y", "em←→en", "a←→a", "um←→un", entre outras, porque as palavras correspondentes coocorrem em segmentos paralelos.
Munidos deste novo conhecimento é possível efectuar um realinhamento mais fino do texto [ver Tabela 2].

[5] É estudado em pormenor na secção 3.3, o comportamento das medidas de semelhança para cada limiar, por isso o limiar a utilizar definitivamente resulta do estudo efectuado posteriormente.

1	A	**	La
2	Cimeira		cumber
3	Ibero-americana	**	Iberoamericana
4	De		de
5	Madrid	**	Madrid
6	Constituiu		ha significado
7	Um	**	un
8	Progresso		
9	Substancial	**	sustancial
10	relativamente a a		avance
11	cimeira realizada		
12	Em	**	en
13			relación a la celebrada
14	O	**	el
15	Ano	**	año
16	Anterior	**	anterior
17	Em	**	en
18	Guadalajara	**	Guadalajara
19	,	**	,
20	em o		(
21	México	**	México
22			)
23	,		,
24	porquanto		pues
25	Se	**	se
26	Aprovaram	**	han aprobado
27	Programas	**	programas
28	Concretos	**	concretos
29	,	**	,
30	Especialmente	**	especialmente
31	Em	**	en
32	O	**	el
33	Âmbito	**	ámbito
34	De		de
35	A	**	la
36	Educação	**	educación
37	E	**	y
38	de a		

39	Cultura	**	cultura
40	,	**	,
41	de a		
42	Saúde	**	salud
43	,	**	,
44	de o		
45	Mundo	**	mundo
46	Indígena	**	indígena
47	E	**	y
48	de o		
49	Meio	**	medio
50	Ambiente	**	ambiente
51	.	**	.

Tabela 2: Realinhamento dos dois textos usando as palavras cognatas

O alinhamento entre os correspondentes textos paralelos PT-EN, é mostrado na tabela seguinte, onde são marcadas com um asterisco as traduções conhecidas (aceites e validadas).

1	The	*	A
2	Latin American		
3	Summit	*	cimeira
4			Ibero-americana
5	held in		de
6	Madrid	*	Madrid
7	achieved considerable		constituiu um
8	Progress	*	progresso
9	compared with the one held		substancial relativamente a a cimeira realizada em o
10	last year	*	ano anterior
11	In	*	em
12	Guadalajara	*	Guadalajara
13	(		, em
14	Mexico	*	O México
15	)		
16	,	*	,
17	with the adoption of specific		porquanto se aprovaram

18	Programmes	*	programas
19			concretos
20	,	*	,
21	Notably		especialmente
22	In	*	em
23	The	*	e
24	Fields		âmbito
25	Of	*	de
26			a
27	Education	*	educação
28	And	*	e
29			de
30	Culture	*	a cultura
31	,		,
32			de a
33	Health	*	saúde
34	,	*	,
35	indigenous peoples		de o mundo indígena
36	And	*	e
37			de
38	The	*	o
39	Environment	*	meio ambiente
40	.	*	.

Tabela 3: Alinhamento dos textos paralelos EN-PT

Assumindo as correspondências estabelecidas nos alinhamentos dos textos paralelos PT-EN, utilizando traduções conhecidas de palavras e expressões, é possível extrair as seguintes traduções para o par EN-ES:"the ←→ la", "Madrid ←→ Madrid", "last year ←→ año anterior", "year ←→ año", "last ←→ anterior", "Mexico ←→ México", "programmes ←→ programas", "notably ←→ especialmente", "in ←→ em", "fields ←→ámbito", "education ←→ educación", "culture ←→ cultura", "health ←→ salud", "environment ←→ medio ambiente", pelo menos. Realinhando cada um dos pares de textos e aplicando o extractor de [Lopes, Gabriel e Aires, José, 2009] conseguir-se-ia extrair os termos "latin american summit←→cimeira ibero-americana←→cumbre iberoamericana" ou "summit←→cimeira←→cumbre" e, provavelmente, para todo o corpus, outras traduções seriam extraídas.

Será ainda, utilizado um filtro da posição relativa dos termos no texto. Este filtro permite eliminar o emparelhamento dos termos cujas coordenadas relativas não se situem na

mesma zona em ambos os textos, o que corresponde à diagonal principal ilustrada pela figura em baixo que foi usada na tese do António Ribeiro [Ribeiro, António, 2002].
Em duas línguas tão similares, a localização no texto de termos que são tradução um do outro não difere muito. Neste caso, qualquer emparelhamento de termos que se desvie até determinado limite da diagonal principal, deverá ser rejeitado.
A utilização deste tipo de filtro evita os falsos amigos, levando a uma maior precisão do algoritmo.

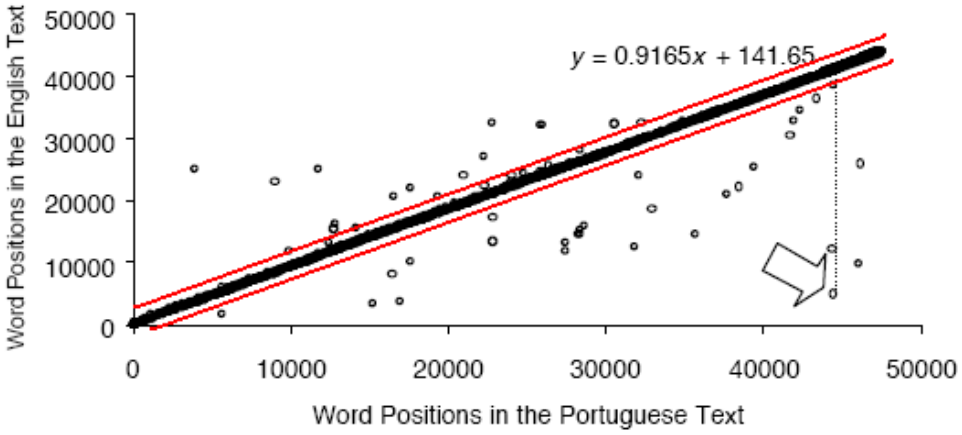

Figura 2: Distribuição das posições de uma palavra e a sua possível tradução

A figura a cima, representa o gráfico da distribuição das posições de uma palavra e a sua possível tradução, relativas a um texto em português e o seu texto paralelo em inglês. O eixo das abcissas marca a posição da palavra no texto em português, enquanto o eixo das ordenadas marca a posição da palavra no texto em inglês [Ribeiro, António, 2002].
Certamente que esta não é a única forma de fazer a correspondência entre termos, porque apesar dos dois idiomas terem semelhanças notórias, têm também diferenças a nível da construção frásica. Deve, no entanto, ter-se em conta a liberdade do tradutor, na escolha de determinados termos em detrimento de outros.
Ainda que, o novo léxico tivesse necessitado de validação humana, conseguiu-se um grande número de entradas válidas que não requeriam grandes correcções, aumentando desta forma os léxicos de EN-ES e de PT-ES, em quantidade e qualidade e diminuindo o tempo de validação requerido. Além disso, as entradas que forem consideradas como incorrectas e marcadas como tal, podem sempre ser utilizadas em conjunto com as que foram consideradas correctas, para treinar um classificador que automaticamente, possa tirar partido do conhecimento que fica implícito nessas classificações, podendo em

particular, ajudar-nos face à filtragem de novas extracções, antes de elas serem passadas a um validador humano [Henderson, J.C., 2003][6].

1.3Principais contribuições

Ao contrário do que é usual em trabalhos sobre extracção de léxicos bilingues via pivotagem, onde existem dados e conhecimento prévio sobre dois pares de línguas X-Y e Y-Z, a extracção de novo conhecimento é feito para o par X-Z, servindo a língua Y de pivô. Partindo do conhecimento validado apenas sobre um único par de línguas (PT-EN), do conhecimento que temos sobre como alinhar textos paralelos a um nível subfrásico [Ribeiro, António, 2002; Gomes, Luís, 2009], e das várias medidas de semelhança entre unidades textuais de tamanhos diversos, foram extraídos dois léxicos bilingues para os dois pares de línguas para os quais não se dispõe de léxico. Foram experimentados diversos limiares de aceitação, tanto para medidas de semelhança como frequências de ocorrência. Procedeu-se à validação dos resultados obtidos, discutindo-os em função das medidas de semelhança utilizadas e dos limiares escolhidos para aceitação das traduções automaticamente extraídas. Enquadrei estes resultados comparando-os com resultados obtidos em trabalho equiparável.

Este trabalho contribui com uma medida de semelhança nova (*Levenshtein*Dice*), que teve consequências positivas relativamente à precisão da extracção de traduções. Conseguiu-se um ganho substancial na construção de léxicos bilingues de alta qualidade, reduzindo o esforço humano nas validações das entradas de cada novo par de línguas. O uso de várias fontes de conhecimento não independentes permitiu restringir grandemente e de forma mais precisa a qualidade e a quantidade do conhecimento extraído.

A principal contribuição centra-se na utilização em simultâneo de três pares de línguas, utilizando apenas um léxico altamente evoluído de um desses pares, para orientar o processo de extracção. São apresentados os valores de precisão e cobertura dos léxicos obtidos [ver secção 3.3].

[6] Há trabalho em desenvolvimento (comunicação pessoal), realizado por Kavitha Mahesh, que aponta nesta direcção.

2. Trabalho relacionado

Este capítulo descreve uma visão geral do trabalho anterior relacionado com o tema desta tese. Em particular, tratar-se-ão dos processos de alinhamento de textos paralelos e de extracção de equivalentes de tradução, que permitem a criação automática de léxicos bilingues, de forma a melhorar o alinhamento de textos subsequentes.

2.1 Alinhamento de textos paralelos

Nesta secção será descrito o método apresentado na dissertação de Mestrado de Luís Gomes [Gomes, Luís, 2009], pois será este o método usado para alinhar os textos que serão objecto do trabalho a desenvolver.

O alinhamento de dois textos paralelos consiste em dividir os dois textos em vários segmentos, tais que o *iésimo* segmento de um texto corresponda ao *iésimo* segmento do outro, isto é, para que sejam tradução um do outro.
Em trabalhos que precederam e serviram de base à referida tese, assume-se que a combinação de várias técnicas (cognatos, palavras homógrafas, léxicos internos e externos) para encontrar correspondências é melhor do que utilizar apenas uma única dessas técnicas.
Ficou provado que, usando apenas um léxico bilingue, cujas entradas tenham sido extraídas automaticamente e validadas antes de serem reutilizadas no processo de realinhamento, consegue-se obter um alinhamento mais preciso do que o que era obtido quando se utilizavam possíveis cognatos extraídos como parte do próprio processo de alinhamento [Bilbao, Darriba et al., 2005; Gomes, Luís, 2009].
O processo de alinhamento, na perspectiva destes autores, ao estabelecer correspondências muito seguras entre traduções de termos, mono ou multi-palavra, divide cada um dos textos paralelos em segmentos que, desejavelmente, deveriam continuar a ser paralelos, i.e. traduções uns dos outros. Observa-se porém que a precisão dos alinhamentos obtidos ao nível destes segmentos, estando longe de ser de 100%, aumenta no entanto com a quantidade e a qualidade do conhecimento disponível sobre a tradução de palavras e multi-palavras. Como é relatado em [Gomes, Luís, 2009], a precisão destes alinhamentos subfrásicos passou de um máximo de 75,5% (quando se utilizaram como alinhadores apenas possíveis cognatos entre PT e EN [Bilbao, Darriba et al., 2005]) para 84,5% quando se utilizou um léxico bilingue com cerca de 60.000 entradas automaticamente extraídas, validadas manualmente, para o mesmo par de línguas.
Em resumo, a separação dos processos de alinhamento e de extracção de cognatos/léxicos, permitiu obter léxicos bilingues mais ricos, porque puderam ser utilizados métodos mais sofisticados para a extracção dos mesmos. Passou a haver maior

controlo sobre o léxico, o que significou que puderam ser retiradas ou corrigidas quaisquer entradas incorrectas. Consequentemente, os alinhamentos em si passaram a requerer menos processamento, tornando todo o processo mais ágil.

O método produz um alinhamento e um mapa de correspondências, onde a diagonal principal [Ribeiro, António, 2002] é utilizada inicialmente como um alinhamento mais grosseiro para obter o mapa de correspondências inicial. Posteriormente, o método itera em duas etapas, que consistem em calcular um alinhamento óptimo a partir do mapa de correspondências e, encontrar novas correspondências utilizando o alinhamento anteriormente obtido como guia. Para refinar os alinhamentos subsequentes, o mapa é incrementado com as novas correspondências e, por esse motivo, o alinhamento obtido em cada iteração é pelo menos tão bom como o da iteração precedente. Este ciclo termina assim que o alinhamento deixa de refinar.

De uma forma geral o algoritmo funciona da seguinte forma:

1. Obter os vectores de ocorrências de equivalentes de tradução a partir dos textos paralelos.
2. Recorrer à diagonal principal como uma medida mais grosseira, gerar o mapa de correspondências entre os pares de ocorrências obtidas no passo anterior.
3. Seleccionar do mapa de ocorrências o alinhamento com maior cobertura.
4. Obter novas correspondências usando como guia o alinhamento obtido no passo 3, adicionando-as também ao mapa.
5. Repetir os passos 3 e 4 até que o alinhamento deixe de refinar, ou seja, até que a cobertura dos alinhamentos pare de aumentar.

Convém referir, que o método de alinhamento proposto por [Gomes, Luís, 2009] é uma evolução da abordagem proposta por [Ribeiro, António, 2002] que diverge claramente da evoluções de propostas de alinhamento [Brown, P.F. et al., 1991; Gale, W.A. et al., 1993; Och, F. et al., 2003] que procuram obter um alinhamento à frase, procedendo-se depois a um alinhamento à palavra, onde cada uma das palavras de cada uma das frases é assumida como podendo ser tradução de qualquer das palavras das frases paralelas, atribuindo a cada par de palavras em cada uma das línguas, uma probabilidade de ser tradução. Ou seja, nesta outra metodologia o alinhamento de palavras e a extracção de traduções de palavras são processos indistinguíveis.

2.2Extracção de equivalentes de tradução

A produção de bases de dados lexicográficas multilingues, como o projecto do *Joint Research Centre* da Comissão Europeia (JRC) ou o projecto *Papillon*, é uma área de grande interesse para a tradução automática, uma vez que as mesmas podem ser utilizadas por tradutores humanos ou por aplicações de processamento de língua natural.
Com o grande aumento de corpora paralelos observado nos últimos anos, o seu processamento automático torna-se absolutamente necessário.
Neste panorama, o alinhamento de textos paralelos torna-se preponderante para a extracção de léxicos multilingues, existindo para o efeito vários métodos. As demais técnicas de alinhamento diferem na abordagem, desde os métodos puramente estatísticos até aos métodos com uma abordagem linguística. Mas a maior divergência entre eles, dá-se ao nível da granularidade dos alinhamentos (parágrafo, frase, termos compostos ou simples). Os métodos mais sofisticados combinam técnicas estatísticas com pistas linguísticas, permitindo alinhamentos de granularidade mais fina.
Nas subsecções seguintes serão descritos alguns métodos de extracção de léxicos.

2.2.1 Equivalentes de Tradução

Entende-se por equivalente de tradução, o par de palavras ou de expressões de línguas distintas que são tradução uma da outra. Um equivalente de tradução pode não corresponder a uma tradução literal, como acontece com a frase ”*Thank you*“ em Inglês e a palavra “*Merci*“ em Francês ou “*counterclockwise*” em inglês e “no sentido contrário ao dos ponteiros do relógio” em portugês.
A extracção de equivalentes de tradução é uma das tarefas mais importantes para a construção automática de léxicos bilingues. Este tipo de recurso linguístico é extremamente útil para os sistemas de tradução automática, pelo que os textos paralelos revelaram ser ideais para obter equivalentes de tradução, pois fornecem o texto original e as suas traduções equivalentes noutras línguas.

Esta secção descreve sucintamente o método de José Aires e Gabriel Lopes [Lopes, Gabriel e Aires, José, 2009], para extracção de traduções de unidades frásicas (sequências não interrompidas de palavras como é o caso de “meio ambiente” que se traduz por “environment”) a partir de corpora paralelos alinhados, usando Suffix Arrays e outras estruturas de dados relacionadas.
Os alinhamentos à unidade frásica permitem uma redução significativa do espaço de pesquisa. Esta abordagem não evita no entanto, os alinhamentos incorrectos resultantes de erros, características específicas das línguas e de tradução incorrecta ou de pouca

qualidade. Contudo é de assinalar que a precisão do alinhamento PT-EN evoluiu de um máximo de 75,46% [Bilbao, Darriba et al., 2005], quando apenas se utilizavam possíveis cognatos como pontos de correspondência, para 84,6% quando se utilizou um léxico validado automaticamente extraído com cerca de 60 000 entradas [Gomes, Luís, 2009].
Com os textos paralelos finamente alinhados, torna-se mais simples efectuar a associação entre os termos de origem e os de destino, desde que o acesso a ambos seja eficiente.
O objectivo é obter para cada termo do texto de origem, o conjunto de segmentos adjacentes que encaixem exactamente no termo de origem, e depois obter os termos adjacentes correspondentes no texto de destino. Uma vez que só interessa nova informação, os equivalentes de tradução compostos por segmentos conhecidos não são tidos em linha e conta. Uma das características que estes autores utilizaram resulta de um número considerável de padrões como acontece na Tabela 4 em que "*Latin American*" aparece a alinhar com nada, seguido de "*summit*" a alinhar com "cimeira" e anotado como conhecido, seguido de "ibero-americana" a alinhar novamente com nada. O aproveitamento deste tipo de padrões para propor possíveis traduções, depois de devidamente filtradas utilizando uma ou várias medidas de semelhança ou uma combinação ponderada dessas medidas, costuma conduzir à extracção de bons equivalentes de tradução. Seria o caso de "latin american summit ←→ cimeira ibero-americana".

1	The	*	A
2	Latin American		
3	Summit	*	cimeira
4			Ibero-americana

Tabela 4: Exemplos de termos sem alinhamento

Claro que o mesmo padrão permite também a extracção de equivalentes errados como "latin american summit ←→ cimeira" ou "summit ←→ cimeira ibero-americana". Mas é para ultrapassar estas contingências que o trabalho apresentado na cadeira de MLKE que Kavitha Mahesh treinou um classificador [Joachims, Thorsten, 1998] utilizando do dicionário bilingue existente as entradas que haviam sido classificadas maioritariamente como correctas e as que haviam sido classificadas como incorrectas. Utilizando vários factores como as probabilidades condicionais (num sentido e no outro) para cada par de expressões classificadas como correctas ou incorrectas bem como a localização relativa e outras *features* conseguiu-se atingir uma precisão de 98% no corpus paralelo utilizado. A partir daí, utilizando o extractor de José Aires [Lopes, Gabriel e Aires, José, 2009] extraíram-se cerca de 190.000 possíveis traduções e procedeu-se à sua classificação. Daí resultara cerca de 40.000 entradas classificadas automaticamente como correctas e 150.000 classificadas como incorrectas. A validação manual dessas entradas é feita a uma

velocidade de cerca de 600 a 1200 entradas por hora apesar de a precisão real não ser igual aos 98% obtidos pelo classificador com o conjunto de teste mas andará acima dos 90% quer para as entradas classificadas automaticamente como correctas ou como incorrectas.
Um equivalente de tradução possui as seguintes três propriedades que são aproveitadas para calcular as medidas de semelhança utilizadas no melhoramento dos resultados:

1. A frequência de emparelhamento, que corresponde ao número de vezes que determinado termo de origem é associado a um termo de destino
2. A frequência do termo de origem
3. A frequência do termo de destino

Uma vez que o método requer um acesso rápido aos termos e às suas frequências, esta solução utiliza *Suffix Arrays* como estrutura de dados para armazenamento e consulta dos vários termos do corpus, já que esta estrutura providencia um acesso rápido aos termos, permitindo calcular eficientemente as suas frequências para utilização nas medidas de classificação usadas.

2.2.2 Geração de bases de dados lexicográficas multilingues a partir de textos paralelos usando recursos endógenos

A abordagem proposta por Giguet Emmanuel e Luquet Pierre-Sylvain [Giguet, Emmanuel et al., 2006] centra-se na obtenção de equivalentes de tradução com múltiplas granularidades.
Sendo esta abordagem endógena, utiliza somente o corpus paralelo multilingue do *Acquis Communautaire* (AC) como *input*, sem utilizar qualquer outro recurso linguístico. Não recorre inclusive, a anotações linguísticas nem à normalização de palavras (nomeadamente à lematização que implicaria a alteração das várias formas verbais para a forma infinitiva, das formas adjectivais para a forma masculina singular, das formas nominais para a forma singular).
O algoritmo lida com um par de línguas de cada vez, recebendo como *input* um bitexto (ou dois textos paralelos) **alinhado à frase**. Note-se a diferença com a metodologia apresentada na secção anterior em que o alinhamento é subfrásico e não à frase.
Cada bitexto é um quadruplo da forma <T1, T2, Fs, C> onde T1 e T2 são textos paralelos, Fs é uma função que reduz T1 e T2 a um conjunto de elementos, respectivamente Fs(T1) e Fs(T2), e C é um subconjunto do produto cartesiano de $\mathrm{Fs(T1)} \times \mathrm{Fs(T2)}$.
A aplicação do método é organizada em duas fases, que se baseiam em duas hipóteses subjacentes.

Hipótese 1: Considere-se um bitexto composto pelos textos T1 e T2. Para determinada sequência S1 que se repete várias vezes em T1 e em termos bem definidos, existem fortes possibilidades de uma sequência S2 que corresponda à tradução de S1, ocorrer nos alinhamentos correspondentes em T2.
Hipótese 2: Considere-se um corpus de bitextos, composto por duas línguas L1 e L2. Não existem garantias da sequência S1, que é repetida em inúmeros textos da língua L1, possua uma única tradução nos textos correspondentes da língua L2.

Para ilustrar a Hipótese 2 temos o caso da palavra inglesa *ticket* que em português pode significar bilhete (cinema) ou multa (trânsito), dependendo do contexto em que se aplicam.

A primeira fase de aplicação do método definido por Giguet Emmanuel e Luquet Pierre-Sylvain [Giguet, Emmanuel et al., 2006], corresponde à análise dos bitextos, sendo aplicada somente a nível do documento, pelo que cada um é trabalhado individualmente.
Esta fase inicia-se com a obtenção das sequências repetidas e as suas frequências, processando os documentos de ambas as línguas L1 e L2 independentemente.
O tratamento da flexão das palavras é também considerado nesta fase. O método consiste em efectuar aproximações à fronteira que delimita o núcleo da palavra e o seu sufixo. Esta é uma variante do método do pico e do plateau [Frakes, W. B. et al., 1992] que é aplicada na detecção de radicais de palavras, e que por isso, olha para as cadeias de caracteres da esquerda para a direita. Este limite marca a posição onde o número de letras que precede um sufixo de comprimento n é superior ao número de letras precedentes de um sufixo de comprimento n-1. Para ilustrar este método, pode observar-se que no primeiro documento em Inglês do corpus AC, "g" é precedido por 4 letras, "ng" por 2 e "ing" por 10. Neste caso pode afirmar-se que "ing" é provavelmente um sufixo. No documento correspondente em Grego, "ά" é precedido por 5 letras, "κά" por 1 e "ικά" por 10. Pode então afirmar-se que "ικά" é possivelmente um sufixo.

As sequências são representadas vectorialmente, para que a detecção das respectivas traduções seja calculada pelo coseno [ver secção 2.3.3] dos vectores associados a elas. Estes vectores obtêm-se armazenando em cada célula o número de ocorrências de determinada sequência no segmento de índice correspondente ao índice da célula em questão. Note-se que aqui o alinhamento é à frase e a célula corresponde aos segmentos ali representados.
Para o alinhamento de cada sequência da língua L1, obtêm-se a relação de tradução entre determinada sequência de L1 e cada sequência de L2 a ser alinhada. Este tipo de relação entre duas sequências é determinado pelo coseno dos seus vectores correspondentes.

A segunda fase compreende o processamento a nível do corpus, filtrando e classificando os alinhamentos.
Determinado termo pode possuir várias traduções de acordo com o domínio do tema ou género do documento em que se encontra.
A filtragem de termos aceita aqueles que foram extraídos em pelo menos dois documentos, ou cuja extracção foi efectuada unicamente num documento, desde que os termos alinhados correspondam à mesma sequência ou ainda, se a frequência dos seus termos for superior a determinado limite. Este limite é proporcional ao inverso do comprimento do termo, desde que existam menos termos complexos repetidos do que termos simples.
Esta abordagem demonstra que é possível contribuir para o processamento de textos em línguas sobre as quais existam poucos recursos linguísticos disponíveis.

2.2.3 Extracção de léxicos bilingues a partir de corpora paralelos e não paralelos

Os léxicos bilingues são auxiliares de tradução de palavras ou frases, extremamente úteis aos tradutores humanos. Pela mesma razão, também são de extrema importância para melhorar a qualidade dos alinhamentos e, consequentemente as traduções, ao adicionar informação fidedigna ao processo de decisão, criando assim alinhamentos com um grau de correcção elevado.
Os algoritmos para obtenção automática de léxicos bilingues a partir de textos paralelos alinhados exploram várias características dos alinhamentos destes textos. Resultante do facto de serem paralelos, para um mesmo tema ou domínio, as palavras possuem o mesmo contexto em ambas as línguas, bem como padrões de utilização muito semelhantes [Fung, Pascale, 1998].

Nesta secção do documento, descreve-se o algoritmo proposto por [Fung, Pascale, 1998], que propõe uma aproximação estatística à extracção de léxicos bilingues a partir de corpora paralelos e não paralelos.
Muitos dos textos paralelos são limpos manualmente para eliminar ruído da tradução. Um corpus diz-se limpo, quando contém um nível mínimo de ruído, como frases de determinado texto que não possuem tradução no texto paralelo, ou quando os limites das frases não estão bem definidos.
A utilização de corpora paralelos limpos para tradução automática sofreu algumas objecções por parte da comunidade científica. A principal objecção deve-se ao facto de que limitar os recursos somente aos textos paralelos limpos ser demasiado restritivo. Por esse motivo, o método apresentado nesta secção permite extrair das palavras, informação

estatística a partir de diferentes tipos de textos, incluindo corpora paralelos com ruído e textos que embora sejam não paralelos, o seu conteúdo incide sobre o mesmo domínio.

A extracção de léxicos bilingues a partir de corpora paralelos explora as seguintes características do texto:

1. As palavras têm na sua maioria uma única conotação por corpus[7]
2. Não existem frases sem correspondência entre textos paralelos
3. As frequências das palavras nas duas línguas são comparáveis
4. As coordenadas relativas das palavras em ambos os textos são comparáveis

A maioria dos textos paralelos são específicos de determinado domínio, pelo que as palavras geralmente possuem o mesmo sentido e são consistentemente traduzidas pelas mesmas palavras. Uma vez que o corpus esteja alinhado à frase, é possível aprender a correspondência das palavras entre os dois textos. Note-se, mais uma vez, que o alinhamento inicial é à frase.
O alinhamento de corpora paralelos, em que os limites das frases não se encontram bem definidos e que não possuem correspondência directa entre frases nas duas línguas, é difícil de obter.
Dkvec é um algoritmo que em vez de utilizar as coordenadas relativas de uma palavra no texto, utiliza o vector de distâncias da palavra e compara-o com o das palavras a traduzir. Um vector de distâncias é um vector cujas células contêm as diferenças posicionais entre duas ocorrências sucessivas de determinada palavra no texto. Baseia-se na noção de que palavras semelhantes não ocorrem exactamente na mesma posição em cada parte do corpus. As distâncias entre instâncias da mesma palavra são similares entre textos paralelos dessas línguas. Os pares de palavras encontrados são então usados para alinhar os textos paralelos em segmentos. A partir do corpus alinhado, as palavras são representadas num vector de posições binário, sendo emparelhadas usando o método de Informação Mútua [ver secção 2.3.5].

Com o aparecimento da Internet, torna-se evidente que os textos não paralelos são muito mais abundantes, actualizados e mais acessíveis que os textos paralelos. Em determinado tempo, existirá maior número de palavras novas nos textos não paralelos que nos textos paralelos, pois a publicação destes últimos, possui um lapso temporal significativo, resultante do processo de tradução. Por este motivo, é tão apetecível o mapeamento

[7] Esta característica aplica-se melhor às línguas tratadas por Fung, o inglês e o chinês, por serem morfologicamente pobres.

bilingue de palavras para a extracção de léxicos a partir de corpora não paralelos, mas comparáveis nas línguas de origem e destino.

Ao contrário dos textos paralelos, que são definidos como textos traduzidos, existe uma grande variedade de não paralelismo nos textos disponíveis para qualquer língua. Este não paralelismo manifesta-se nas diferenças de autor, domínio, no período temporal e linguagem empregue.
À medida que o não paralelismo dos textos aumenta, mais difícil se torna encontrar padrões estatísticos nos termos. O não paralelismo evidencia as seguintes características:

1. As palavras possuem múltiplos significados por corpus
2. As palavras têm múltiplas traduções por corpus
3. As traduções de algumas frases podem não existir no documento alvo
4. A frequência da ocorrência das palavras não é comparável
5. As coordenadas relativas das palavras em ambos os textos, não são comparáveis

Os textos comparáveis possuem também as seguintes características:

1. Para um mesmo tópico, as palavras possuem contextos comparáveis entre línguas
2. Palavras do mesmo domínio, em que os textos onde se encontram sejam do mesmo período temporal, possuem padrões de utilização comparáveis

O método ***Convec*** encontra candidatos para tradução de novas palavras em corpus comparáveis, usando informação do contexto da palavra para encontrar a sua correspondente na outra língua. O objectivo deste algoritmo é aumentar um dicionário com novas palavras, melhorando a qualidade das traduções efectuadas pelos sistemas de tradução automática que o utilizem.
Este algoritmo encontra pares de palavras em línguas diferentes a partir de texto comparável e não paralelo, usando uma abordagem de Recuperação de informação (*Information Retrieval*).
O algoritmo *Convec* extrai os contextos (conjuntos de cinco a dez palavras em redor dos termos a analisar [ver Figura 3]) de palavras desconhecidas na língua de origem e trata-as como uma *query*. Analisa então os contextos das traduções candidatas na língua de destino que melhor se adaptem à *query*.

Figura 3: Exemplo de contexto da palavra *flu* retirado de artigos de jornais ingleses

O algoritmo funciona da seguinte forma:

1. Obter os vectores de contexto [ver Figura 3] de todas as palavras desconhecidas W_e na língua de origem.
2. Obter os vectores de contexto das traduções candidatas W_c na língua de destino.
3. Para todos os vectores W_e, W_c calcular a medida $similarity(W_e, W_c)$.
4. Ordenar os vectores segundo esta medida de semelhança (descrita seguidamente).
5. Escolher o vector W_c com a classificação maior como candidato de tradução de W_e.
6. Escolher o par (W_e, W_c) com maior classificação como a nova entrada no dicionário bilingue.

Para obter a medida de semelhança é necessário calcular a frequência dos termos (TF), que consiste na contagem da frequência da ocorrência das palavras do contexto do termo com tradução desconhecida.
Para decidir que dois contextos são de facto semelhantes, é necessário encontrar um mapeamento entre as palavras dos contextos, o que é conseguido utilizando um dicionário, significando assim que as palavras existentes no contexto são todas conhecidas.
Antes de se proceder ao cálculo da medida de semelhança, ainda é necessário obter a frequência inversa do documento (IDF – Inverse Document Frequency – inverso da frequência dos documentos onde ocorre o contexto). Esta medida contabiliza a frequência da ocorrência dos contextos de palavras no corpus inteiro.

O IDF utilizado pelo autor, é dado por:

$$IDF = \log \frac{N_{\max}}{N_i} + 1$$

Onde $N_{\max}$ é a frequência máxima de cada palavra no corpus e N_i o número total de ocorrências da palavra de índice i no corpus. Esta fórmula é modificada relativamente à fórmula usual de IDF.

A i-ésima dimensão do vector de contexto de determinada palavra é dada por $W_i = TF_i \times IDF_i$, que é zero caso a palavra não apareça no contexto.
Para localizar os candidatos para a tradução, é necessário comparar o vector de contexto da palavra desconhecida com os vectores de contexto de todas as palavras dos textos da língua de destino. Neste algoritmo é utilizada uma variante da medida Coseno.

$$S(W_c, W_e) = \frac{\sum (W_{ic} \times W_{ie})}{\sqrt{\sum W_{ic}^2 \times W_{ie}^2}}$$

Onde $W_{ic} = TF_{ic} \times IDF_i$ e $W_{ie} = TF_{ie} \times IDF_i$

Este método pode efectivamente ser usado para extrair léxicos de corpora comparáveis e não paralelos, no sentido de aumentar dicionários baseados em léxicos extraídos de textos paralelos ou criados por humanos.

Pablo Gamallo e José Pichel Campos [Gamallo, Pablo et al., 2005] usam também, corpora não paralelos e etiquetadores morfo-sintácticos como auxiliares para a extracção de traduções de palavras. Mas, sobre este outro trabalho não se entra em detalhe.

2.2.4 Indução de Léxicos Usando Línguas Pivô

Gideon S. Mann e David Yarowsky [Gideon, Mann et al., 2001] apresentaram um método de indução de léxicos que permite relacionar cognatos de pares de línguas através de uma língua pivô.
Os léxicos bilingues entre línguas da mesma família são induzidos utilizando modelos probabilísticos de distância entre cognatos de textos paralelos. Enquanto os léxicos para tradução entre pares de línguas de raízes distintas, são gerados por uma combinação destes modelos de tradução intra-familiar e, um ou mais dicionários on-line para línguas com bases diferentes.

Obteve-se até 95% de precisão no vocabulário de destino, permitindo desta forma, que partes substanciais dos léxicos possam ser geradas com precisão, para idiomas que não possuam dicionários bilingues ou corpora paralelos.

Palavras cognatas são definidas como um par de tradução, onde duas palavras de línguas distintas partilham o significado e a raiz de que são formadas. (Ex: "*neveu*" em Francês e "*nephew*" em Inglês). É claro que nem todas as traduções são cognatos, e em alguns casos, apesar de partilharem a mesma base ou estarem historicamente relacionadas, podem ser de difícil resolução para o modelo. (Ex: "*père*" em Francês, "*father*" em Inglês). Quanto mais semelhantes são duas línguas, maior o número de palavras cognatas partilhadas entre elas.
Foi mostrado que línguas da mesma família são próximas o suficiente, de modo a que os pares de cognatos entre duas línguas são comuns, e porções significativas do léxico podem ser induzidas com alta precisão.

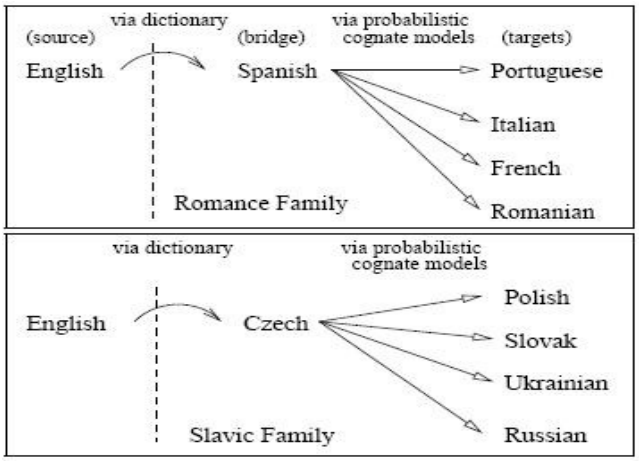

Figura 4: Indução de Léxicos através de línguas pivô

Para ligar línguas com raízes distintas, foi usado um modelo de dois passos através de línguas pivô (como mostrado na figura anterior). Dicionários on-line entre a língua de origem e outra língua representativa de uma família linguística, podem ser combinados com modelos baseados em cognatos, aplicados a línguas da mesma família para produzir léxicos entre a língua origem e todas as restantes pertencentes à família da língua pivô.
O desempenho destes modelos pode ser melhorado utilizando múltiplas línguas pivô, aumentando a cobertura e precisão.

O algoritmo de indução proposto baseia-se num método de determinação da distância (Ex: distância de *Levenshtein*) entre duas palavras de línguas distintas. Esta distância deve ser baixa para os pares cognatos, e alta para os pares não cognatos.

Formalmente, temos:
Dadas duas línguas S e T, onde *cognato* é uma função que indica que um par é cognato, uma boa função de distância D: $S \times T \rightarrow R$ é uma tal que:

$\forall s \in S, \forall t_c, t \in T$:

$\text{Se cognato}(s, t_c) \wedge \text{não Cognato}(s, t)$

$\text{Então } D(s, t_c) < D(s, t)$

Dada tal distância, podemos aplicá-la na criação de novas traduções para as línguas através do mapeamento de cada palavra da língua origem com as da língua destino mais próxima (relativamente à distância D).
Formalmente:

$\forall s \in S, escolher, \hat{t} \in T : \hat{t} = \arg\min D(s,t), t \in T$

Foram testadas três diferentes funções de distância: distância de *Levenshtein*, uma função de custo obtida usando transdutores estocásticos e uma outra obtida por meio de um modelo de *Markov*. Existem diferenças significativas entre a função de distância de *Levenshtein* e os dois métodos probabilísticos: o primeiro é uma métrica estática que não exige qualquer treino, enquanto as posteriores são métricas adaptativas que devem ser treinadas para um determinado conjunto de dados.
O modelo de *Markov* usado é um modelo baseado na forma fonética da palavra. A soma das probabilidades de todas as possíveis sequências de edição é igual à unidade. Ao contrário do modelo dos transdutores estocásticos, as operações atómicas de edição de cada carácter também somam um.
Claramente, estes métodos não são projectados para descobrir os pares de traduções sem relacionamento de forma entre ambos. Contudo, são aplicáveis nas traduções com semelhanças ortográficas ou fonéticas. Estritamente para os fins de obter este vocabulário da língua de destino, um par de tradução é assumido como sendo um par cognato, se a distância de *Levenshtein* for inferior a 3. Este limite arbitrário evita a necessidade de fazer juízos linguísticos sobre as relações cognatas, mas parece identificar um útil, subconjunto do vocabulário de destino com poucos falsos positivos.

Em primeira instância, estes métodos foram testados ao obter léxicos para línguas da mesma família (línguas Românicas), pelo que o respectivo algoritmo é descrito em seguida.

Dado um dicionário entre as línguas S e T:

a) Seleccionar 100 pares de palavras para testar.
b) Para as medidas adaptativas, as quais requerem treino, seleccionar como dados de treino, os hipotéticos pares de cognatos (aqueles com distância inferior a 3) dos pares restantes de palavras. O algoritmo será treinado com estes pares.

c) Para cada palavra no idioma de origem escolher a palavra mais próxima (relativamente à função de distância) na língua de destino da lista de 100 pares.
d) Um possível par de tradução está correcto se coincidir com a tradução dada no dicionário de referência, de outra forma será marcado como incorrecto. (Assume-se existir apenas uma tradução por palavra. Estamos a investigar modelos que admitem várias traduções de cada palavra.)

	Model	Spanish-Portuguese		French-Portuguese	
		cognate vocab (68%)	full vocab	cognate vocab (39%)	full vocab
L	Levenshtein	92.3	67.9	66.4	32.0
H	Hidden Markov Model	82.2	58.6	62.7	30.0
S	Stochastic Transducer	92.3	67.1	78.6	38.5
L-V	Levenshtein w/vowel sensitive distance	91.9	67.9	68.4	33.8
L-A	Levenshtein w/learned weights (pan-family)	92.9	67.9	80.1	40.5
L-S	Levenshtein w/learned weights (single language)	**94.7**	69.8	**84.3**	42.3

Figura 5: Resultados dos testes efectuados às diversas medidas de distância consideradas.

A tabela anterior mostra os resultados para as diferentes funções de distância usadas para a obtenção das traduções dos pares Espanhol-Português e Francês-Português.
As métricas descritas nas três primeiras linhas, são a distância de *Levenshtein* (L), o modelo de *Hidden Markov* (H), e o transdutor estocástico (S). Os outros três métodos são variantes de distância *Levenshtein* onde os custos para as operações de edição foram modificados. No L-V, as operações de substituição entre as vogais são alteradas de 1 para 0,5.
As restantes variantes adaptativas, L-S e L-A, são mostradas nas duas últimas linhas. Os pesos destes dois sistemas foram produzidos pela filtragem das probabilidades obtidas a partir do transdutor estocástico em três classes de peso 0,5, 0,75 e 1.
Para L-S, a matriz de custos foi treinada em separado para cada par de línguas, enquanto que para L-A, foi treinada colectivamente sobre todas as línguas Românicas consideradas.
Como pode ser observado na Tabela, a distância de *Levenshtein* obtém excelentes resultados. A adaptação dinâmica através da métrica dos transdutores estocásticos (S) também dá um incremento notável para o par Francês-Português, aumentando a precisão dos cognatos, mas oferece pouca melhora no par Espanhol-Português.
Além disso, empiricamente sugere que o melhor método é conseguido através da aprendizagem dos pesos recorrendo aos transdutores estocásticos e, em seguida, usar estes pesos no método L-S.

2.2.5 Indução de Léxicos Usando Diversas medidas de Semelhança e Línguas Pivô

O crescimento explosivo da internet nos últimos oito anos produziu um aumento correspondente no número de línguas do mundo para as quais já se encontram disponíveis inúmeros textos on-line.
O trabalho desenvolvido por Charles Shafer e David Yarowsky [Shafer, Charles et al., 2002] apresenta um método de indução de léxicos entre duas línguas distantes, sem necessidade de quaisquer corpora paralelos bilingues. O algoritmo combina com a similaridade de ocorrência temporal entre as datas em corpora das notícias, similaridade do contexto entre línguas, a distância de *Levenshtein* ponderada, a frequência relativa e medidas de similaridade "*burstiness*". Estas medidas de similaridade são integradas com o conceito de língua pivô sob um robusto método de combinação de classificadores para ambas as famílias de línguas Eslavas e do Norte da Índia.

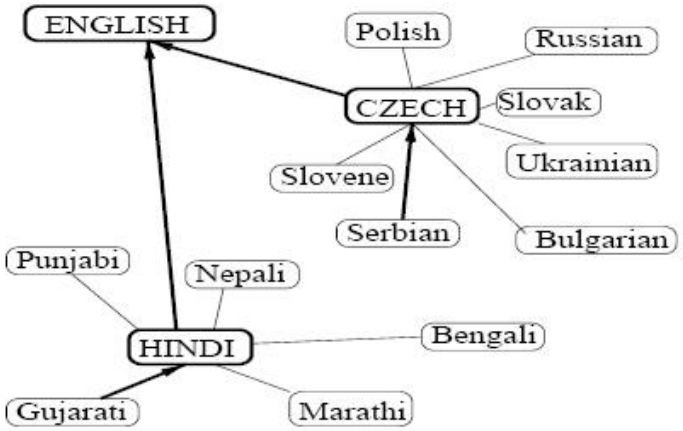

Figura 6: Ligações entre línguas Eslavas e do Norte da Índia usando Inglês como língua pivô.

O objectivo deste algoritmo é o de obter léxicos utilizando recursos que estão disponíveis na internet, sem qualquer custo monetário. Nenhum dicionário é necessário entre o Inglês e a língua de destino, no entanto, é necessário um dicionário de dimensão importante entre a língua pivô e o Inglês. O trabalho com a língua Sérvia envolveu o uso de um dicionário Inglês-Checo inicialmente contendo cerca de 171 000 pares Inglês-Checo, incluindo 54 000 de tipos de palavras unicamente Checas e 43 000 tipos de palavras unicamente Inglesas. O dicionário Hindi-Inglês continha cerca de 74 000 pares. Os vocabulários de Sérvio-Gujarati usados foram construídos pela extracção de todos os tipos de palavras do respectivo corpus, filtrando as de baixa frequência (já que modelos de similaridade usados, carecem de estatísticas confiáveis) e palavras muito curtas (em

experiências preliminares, a utilização da distância entre palavras para propor candidatos para cognatos de palavras muito curtas foi considerado pouco fiável, de modo que palavras com comprimento inferior a 5 caracteres foram excluídas).
No trabalho realizado, excluiu-se numa primeira fase palavras de tamanhos inferiores a 3 caracteres. Numa segunda fase, usando pivotagem, optou-se por considerar estas palavras, pois verificou-se que muitas foram correctamente traduzidas.

O algoritmo apresentado é baseado na nova combinação das seguintes 4 categorias de modelos de similaridade: semelhança de palavras, semelhança de contexto, semelhança da data de distribuição, e semelhança da frequência de palavras e estatísticas "*burstiness*". Três destas 4 categorias são subdivididas em medidas de semelhança individuais, num total de 8: distância de *Levenshtein* ponderada, abrangência do contexto, semelhanças baseadas na data das notícias locais e mundiais, a frequência relativa, "*burstiness*" e frequência inversa de documentos (IDF)[8].
O conjunto inicial de pares de tradução candidatos é gerado considerando todas as palavras da língua origem, com uma baixa distância ponderada, face às entradas no dicionário da língua pivô e Inglês. Os pares candidatos resultantes são então filtrados e classificados pelas medidas de semelhança descritas abaixo.

Semelhança de *Levenshtein* Ponderada

Na primeira iteração, a distância de *Levenshtein* utiliza uma matriz independente da linguagem, que atribui a $\mathrm{dist}(\mathrm{Vogal}\,1, \mathrm{Vogal}\,2)$ e a outras operações sobre vogais, metade do custo das operações equivalentes sobre consoantes (substituições, inserções e exclusões). No início da segunda iteração do modelo, a matriz de distância dos caracteres é novamente estimada utilizando o output da primeira iteração como dados de treino [Gideon, Mann et al., 2001]. Para cada um dos primeiros 2000 pares de tradução de palavras Sérvio-Inglês propostos após a primeira iteração, as palavras Sérvias e as palavras pivôs Checas com a menor distância, são utilizadas como um par no conjunto dos dados de treino para melhorar as ponderações.

Semelhança de Contexto

Para obtermos uma medida de semelhança de contexto, são gerados vectores de conjuntos de palavras para ambas as janelas envolventes de cada palavra no corpus

[8] Ver em pormenor na secção "Combinação de Medidas de Semelhança" da página 32.

[janelas largas (raio de 10) e estreitas (raio 1)], tanto para o Inglês como para a língua de origem (Sérvio, Gujarati). Os vectores da língua de origem são depois traduzidos, utilizando o léxico actual de tradução para inglês, o qual, de momento, ainda apresenta algum ruído. O léxico inicial é gerado a partir do dicionário Checo-Inglês, processando o conjunto de pares de palavras Sérvio-Checo com valores da medida de distância baixos, e tratando a expansão dos pares de palavras resultantes [Sérvio-(via Checo)-Inglês], como um espaço inicial de pares de palavras com ruído. As iterações subsequentes utilizam os léxicos induzidos na iteração de treino anterior.
Esta abordagem distingue-se pelo facto de não utilizar léxicos da língua de ensaio de/para qualquer outra língua, tornando-o adequado para línguas de baixa densidade.

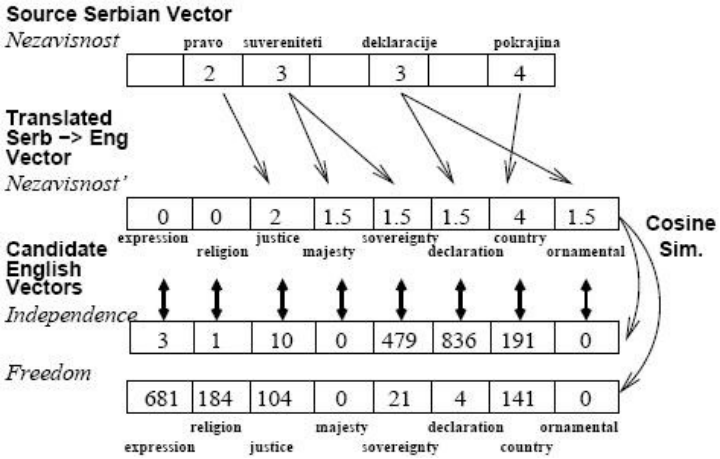

Figura 7: Ilustração do modelo de projecção do co-seno

A figura acima ilustra o modelo de projecção do coseno, comparando os vectores de contexto Sérvios para a palavra a testar *nezavinost* com duas traduções candidatas para Inglês baseadas no modelo de tradução da iteração anterior. A tradução correcta de *nezavinost* (*independence*) apresenta maior similaridade do coseno com o vector da palavra *nezavinost'* do que a alternativa concorrente *freedom*.

Semelhança da Distribuição de Datas

Uma das principais vantagens da utilização de dados de notícias como corpus, deve-se ao facto dos eventos mundiais e regionais (como acidentes de avião, terramotos, golpes de estado, assassinatos, etc.) tenderem a ser relatados em paralelo em vários idiomas e em

datas razoavelmente próximas (geralmente não mais que um dia de desfasamento). Desta forma, ambos os termos Sérvios e Ingleses podem ser representados como vectores de frequência independentes do idioma, ordenados por data ao longo de uma janela temporal de vários anos. Foram compiladas distribuições das datas para cada palavra Inglesa usando como fonte, notícias a nível mundial (todas as notícias Inglesas datadas) e a nível local (notícias da Sérvia em Inglês).

O exemplo em baixo, mostra graficamente como um hipotético par de tradução (correcta) *nezavisnost-independence* tem maior sincronismo na sua distribuição de datas e, consequentemente, uma maior pontuação de semelhança, do que um par concorrente *nezavisnost-freedom* incorrecto, que tem maior classificação pela medida de semelhança *Levenshtein* ponderada, mas é o menor no total das medidas de semelhança, em parte devido à contribuição da semelhança por datas.

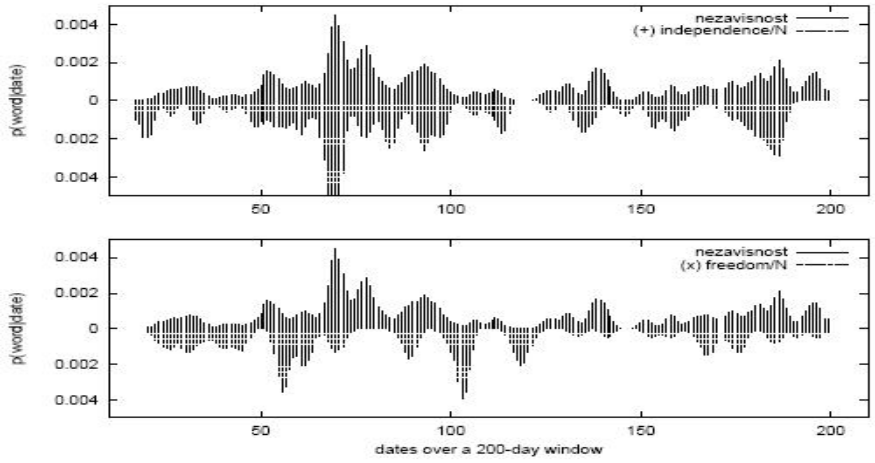

Figura 8: Comparação das distribuições de datas relativas para dois pares de traduções

A figura acima ilustra a comparação das distribuições de datas relativas para o par de tradução correcto *nezavisnost-independence* (similaridade = 0.74) e o par incorrecto *nezavisnost-freedom* (similaridade = 0.42). Em ambos os casos a probabilidade da palavra Sérvia está no eixo das ordenadas positivo e a Inglesa no eixo negativo.

Frequência Relativa

Em média, uma palavra e sua tradução são pouco susceptíveis de ter frequências relativas (RF) semelhantes nos corpora de suas respectivas línguas. Porque o uso polissémico dos

termos de uma língua pode dobrar ou triplicar a sua frequência base, pelo que é espectável verificarem-se ligeiras variações na frequência. No entanto, esta medida é muito útil para descartar a hipótese de pares que exibam diferenças substanciais na frequência relativa. Um simples rácio dos registos de frequências é suficiente, verificando-se algum melhoramento no modelo baseado em pontuações.

$$RFScore = MIN\left[\frac{\log(RF_1)}{\log(RF_2)}, \frac{\log(RF_2)}{\log(RF_1)}\right]$$

EW	RF(EW) ($\times 10^{-7}$)	RF(hvaliti) ($\times 10^{-7}$)	$RFScore_i$
bless/V	64	62	0.998
laud/V	49	62	**0.980**
calibre/N	13	62	0.887
quarter/V	3	62	0.795
class/N	989	62	0.770

Figura 9: Frequência relativa (FR) para a palavra Sérvia *hvaliti*

A tabela representada na figura acima mostra a frequência relativa (FR) para a palavra Sérvia *hvaliti*. A sua correcta tradução (em negrito) pontua valores mais altos do que as restantes alternativas tais como o *calibre*/N e *class*/N. Embora ultrapassem os resultados do termo *laud*/V na semelhança de *Levenshtein* ponderada, as suas frequências relativas observadas, 13 e 989 são significativamente inferiores e superiores (respectivamente) do que o valor de 62 para a tradução do termo *hvaliti*.

Semelhança "*Burstiness*" e Frequência Inversa de Documentos

Church e Gale [Church, K. W. et al., 1995] descreveram várias medidas relacionadas com a tendência das palavras para o contágio da sua distribuição, como ilustrado na Figura 10. Os autores incluem a medida de adaptação $P_{21}(w)$ e a Frequência Inversa de Documentos (IDF). Onde $P_{21}(w)$ é dada por $P(f_w \geq 2 \mid f_w > 1)$ que é a probabilidade da frequência da palavra *w* ser superior ou igual a dois sabendo que a frequência é superior a um e o IDF é usado como uma das medidas de semelhança.

Dada a grande variabilidade de tamanhos de documentos no corpus, também é definida e utilizada uma medida de variação β sobre uma janela móvel de H=200 palavras:

$$\beta = \frac{P(W_i = W \mid W \in \{W_{i-1},...,W_{i-H}\})}{1-(1-P(w))^H}$$

$$\beta Match_i = MIN\left[\frac{\beta_1}{\beta_2}, \frac{\beta_2}{\beta_1}\right]$$

Sendo $\beta Match_i$ a medida de "*burstiness*".

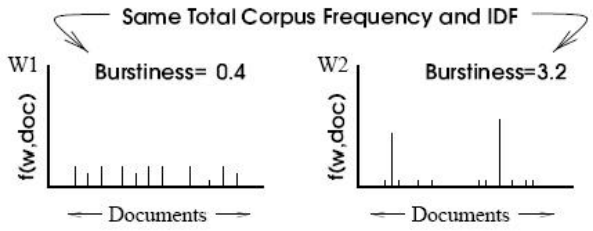

Figura 10: Ilustração da medida "*burstiness*

Ambas as palavras (W1 e W2) têm a mesma frequência no corpus total e o mesmo valor de IDF, mas diferem substancialmente no valor de "*burstiness*".

Combinação de Medidas de Semelhança

A distância de *Levenshtein* ponderada propõe inicialmente um conjunto de pares de tradução candidatos. Para cada par deste conjunto, os valores de semelhança são calculados e as seguintes 8 variantes de medidas de semelhança são utilizadas:

1. A distância de *Levenshtein* ponderada (convertido para uma semelhança, ou seja, uma função crescente de parentesco)
2. Medida de contexto abrangente (10 palavras de raio)
3. Medida de contexto reduzido (raio 1)
4. Distribuição das datas das notícias locais
5. Distribuição das datas das notícias mundiais
6. Frequência relativa (RF)
7. Frequência inversa de documentos (IDF)
8. Medida de semelhança "*burstiness*" (β)

Estes modelos individuais são integrados numa única função de semelhança segundo o seguinte procedimento. Para cada palavra S_a no vocabulário Sérvio (para o idioma Gujarati, sobre o qual não foi realizada normalização, a Etapa 1 é omitida.):

1. *Part-of-Speech (POS) Consistency*: Ao classificar os pares de tradução, foi imposto um favorecimento das partes do discurso compatíveis (substantivo, verbo, adjectivo). A cada palavra Sérvia é atribuído um POS através da análise morfológica, e a cada candidato de tradução Inglês com um POS que não corresponde é dada uma penalização de pontuação suficiente para classificá-los abaixo dos POS dos candidatos compatíveis, mas sem exclui-los (dado que podem ocorrer eventuais erros na atribuição dos POS)
2. *Ranking*: Para cada medida de semelhança S, os candidatos Ingleses são classificados em ordem decrescente de pontuação de similaridade. Às N palavras Inglesas nesta lista ordenada são atribuídos valores, iniciando em 0 para o primeiro item da lista, até N-1, para último item. A cada palavra Inglesa E_b, com o valor C é atribuída uma classificação $R_{norm}(S_a, E_b, S) = \frac{C}{N}$. Onde existem valores de semelhança empatados com as posições i, j na lista, a cada uma das palavras é atribuída uma classificação $R_{norm}(S_a, E_b, S) = \frac{(C_i + C_j)}{N}$.
3. *Scoring*: Cada modelo de semelhança $S_1 \ldots S_8$ tem um peso associado ($\lambda_1 \ldots \lambda_8$) (ver figura em baixo). Para cada palavra inglesa E_b, a pontuação é calculada da seguinte forma:

$$R(S_a, E_b) = \sum_{m:1..8} \lambda_m \cdot R_{norm}(S_a, E_b, S_m)$$

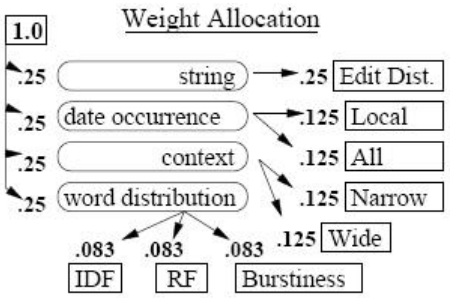

Figura 11: Atribuição de pesos

A figura acima ilustra a atribuição de pesos para a combinação de modelos de baseados em classificações. Como demonstrado, o regime de atribuição de pesos divide em partes iguais, por classe de modelo de semelhança, e executa uma outra divisão igual de pesos para os vários modelos dentro de uma classe.

2.3Medidas de Semelhança

Na criação de modelos de tradução, as medidas de semelhança são utilizadas na fase de estabelecimento de correspondência entre termos, gerando um valor que quantifica o grau de semelhança entre dois termos de línguas distintas. Duas palavras são provavelmente semelhantes, caso o resultado da medida de semelhança aplicada se encontre dentro de determinado limite.
Neste capítulo descrever-se-ão algumas das medidas de semelhança mais conhecidas.

2.3.1 Distância de Levenshtein

Esta função é uma métrica usada na medição da quantidade de diferenças entre dois termos.
A distância de *Levenshtein* entre dois termos é definida como o número mínimo de operações de edição necessárias para transformar um termo noutro, sendo permitidas operações de inserção, remoção ou substituição de um único carácter.
Como as operações são todas de custo unitário, o valor obtido pela medida aquando da comparação de dois termos, é sempre pelo menos a diferença entre os tamanhos dos dois termos e no máximo, o comprimento da sequência mais longa. É zero quando os termos são idênticos.

A título de exemplo, para as palavras “Parlamento” em português e “Parliament” em inglês são necessárias duas operações sobre a palavra inglesa: remoção do “i” e inserção do “o” no final da palavra, obtendo o seguinte resultado.

Levenshtein(“Parlamento”, “Parliament”) = 2

Na introdução a esta dissertação, foi esta a medida usada para a determinação de possíveis cognatos entre palavras portuguesas e espanholas.

2.3.2 Medida de Levenshtein Normalizado

Nesta secção descrevo uma das medidas que utilizei no trabalho experimental descrito nesta tese para identificar palavras cognatas.
Para podermos delimitar os resultados da função de comparação, não será utilizada directamente a distância de *Levenshtein* que é uma medida de dissemelhança, mas sim uma medida de semelhança dada pela fórmula seguinte:

$$Comp(X_n, Y_m) = 1 - \frac{Lev(X_n, Y_m)}{Max(n,m)}$$

Esta função é limitada superiormente por 1 e inferiormente por 0. Sendo X_n uma palavra de comprimento n, Y_m outra palavra de comprimento m, $Lev(X_n, Y_m)$ a distância de *Levenshtein* entre as palavras X_n e Y_m , e $Max(n,m)$ o tamanho da maior palavra.
Interpretando esta fórmula, pode afirmar-se que duas palavras são tanto mais semelhantes, quanto mais próximo de 1 for o resultado da função de comparação.
Dado um termo português (X) e outro espanhol (Y), X é tradução de Y se o resultado de $Comp(X,Y)$ for superior a determinado limite próximo de 1. Este limite será obtido por experimentação. Note-se que esta medida se torna independente do tamanho das palavras comparadas (varia entre 0 e 1).

2.3.3 Coseno

A semelhança do coseno é uma medida de similaridade entre dois vectores, onde é encontrado o coseno do ângulo entre eles.

$$\cos(x, y) = \frac{\sum x.y}{\sqrt{\sum x^2} \times \sqrt{\sum y^2}}$$

O coseno é obtido pela divisão do produto escalar de dois vectores pelo produto das suas normas. Podemos observar que o coseno varia de zero a um, à semelhança das coordenadas dos vectores. As sequências escolhidas para o alinhamento são as que obtiverem maiores valores de coseno. Apesar disso, não se considera determinado alinhamento caso o coseno seja inferior a determinado limite.

2.3.4 Dice

A medida de *Dice* [Salton, Gerard et al., 1983] é uma medida de semelhança entre dois termos X e Y.

$$dice(X,Y) = \frac{2 \cdot f(X,Y)}{f(X) + f(Y)}$$

Onde $f(X,Y)$ representa a frequência da coocorrência dos termos X e Y nos mesmos segmentos e $f(X)$ e $f(Y)$ representam, respectivamente, a frequência de ocorrência de X e de Y nos respectivos textos (ou corpora) alinhados. O valor do coeficiente de *Dice* varia pois entre 0 e 1.

2.3.5 Informação Mútua Específica

Este método mede o grau de relação que existe entre duas palavras. Por exemplo, se a palavra "*constituição*" é frequentemente encontrada perto da palavra "*europeia*", elas podem ter um alto índice de informação mútua. No caso de traduções, a correlação é estabelecida entre palavras ou multi-palavras de línguas diferentes.

É calculado da seguinte forma [Oakes, Michael, 1998]:

$$I(X,Y) = \log_2 \frac{P(X \mid Y)}{P(X)}$$

Sendo que P(X | Y) indica a probabilidade condicional da ocorrência da palavra X dado que a palavra Y ocorreu. P(X) indica a probabilidade da palavra X ocorrer no texto fonte.

2.3.6 Probabilidade Condicional

A medida de probabilidade condicional usada no contexto da extracção de léxicos, mede a expectativa da ocorrência de uma palavra X do texto fonte, sabendo que a palavra Y do texto de destino também ocorre.

$$P(X \mid Y) \approx \frac{f(X,Y)}{f(Y)}$$

Onde $f(X,Y)$ representa a frequência da coocorrência das duas palavras X e Y e, $f(Y)$ a frequência de ocorrência de Y nos textos de uma dada língua.

2.3.7 Qui-Quadrado

O qui-quadrado χ^2 [Smadja, Frank et al., 1996] é uma medida que se baseia num método probabilístico que interpreta um evento num conjunto de documentos. Esta medida testa a hipótese de aceitação ou rejeição das palavras X e Y, de textos paralelos de línguas distintas, co-ocorrerem consistentemente ou por acaso.

É definida como $\chi^2(X,Y) = \frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)}$

Onde:
A – representa o número de vezes que os termos X e Y co-ocorrem em frases paralelas;
B – representa o número de vezes que o termo X ocorre sem o termo Y em frases paralelas;
C – representa o número de vezes que o termo Y ocorre sem o termo X em frases paralelas;
D – representa o número de vezes em que nenhuns dos termos ocorrem em frases paralelas;
N – representa o número total de frases.

2.3.8 Coeficiente de Jaccard

O coeficiente de *Jaccard* é dado pela fórmula [Salton, Gerard et al., 1983]:

$$J(X,Y) = \frac{f(X,Y)}{f(X) + f(Y) - f(X,Y)}$$

A medida de similaridade de *Jaccard* é aplicada a conjuntos de amostras, e é definida como o tamanho da intersecção dividido pelo tamanho da união de todas as amostras.

2.3.9 Outras Medidas

Neste documento já foram encontradas outras medidas que também se utilizaram na tarefa de extracção de traduções de palavras. Para uma lista mais completa ver [Ribeiro, António, 2002] e [Henderson, J.C., 2003].

Recordando o que já foi dito na introdução nesta dissertação, foi utilizada a distância de *Levenshtein* Modificada [ver secção 2.3.2] para a detecção de possíveis cognatos. Como esta medida só analisa a grafia dos termos, não permitindo distinguir as inúmeras situações de traduções de termos parónimos[9], foram obtidas algumas traduções erradas classificadas com valores elevados de *Levenshtein* (próximos de 1). Pode-se excluir algumas destas traduções usando as frequências dos termos no corpus, recorrendo para isso à medida de *Dice* [ver secção 2.3.4]. Para trabalharmos com ambas as métricas em simultâneo será usada uma medida que resulta da multiplicação das duas (*Levenshtein* x *Dice*). Será feito um estudo comparativo destas três medidas que permitirá não só justificar o limiar de aceitação das traduções (ex: $\text{Levenshtein} \times \text{Dice} \geq 0{,}5$), como ilustrar a precisão das medidas face ao limiar seleccionado.

[9] Parónimo, diz-se do vocábulo que tem a mesma origem que outro, que tem o mesmo começo ou a mesma terminação ou um som semelhante e tem significado diferente, in Dicionário Priberam da Língua Portuguesa [em linha], 2009

3. Trabalho Realizado e Análise de Resultados

Este capítulo descreve o trabalho realizado no decorrer desta tese, apresentando um conjunto de metodologias para extracção automática de léxicos bilingues, com recurso a medidas de semelhança entre termos [ver secção 2.3.9] e pivotagem com utilização simultânea de três pares de línguas. Serão analisados os resultados e estatísticas resultantes de aplicação destas técnicas ao corpus seleccionado para o efeito.
A primeira secção introduz o corpus usado indicando estatísticas sobre as suas dimensões e composição. A segunda secção descreve efectivamente o algoritmo concebido ilustrando o seu funcionamento com exemplos extraídos dos textos trabalhados. As duas secções subsequentes tratam da precisão das várias metodologias usadas, analisando os resultados face aos objectivos propostos. A quinta e última secção compara os resultados desta tese com o trabalho de outros autores, permitindo identificar algumas das contribuições desta dissertação para a área de investigação em causa.

3.1 Corpus Usado

Para a obtenção de resultados e demonstração dos algoritmos apresentados nesta dissertação de mestrado foram utilizados alguns textos sobre legislação em vigor da comissão europeia, do capítulo 16.10 sobre Ciência.

Dado que cada tradução de termos extraída, utilizando algumas das medidas de semelhança apresentadas na secção 2.3 ou usando pivotagem, necessita de validação, indicando se a correspondência entre termos extraídos é válida ou inválida, utilizou-se somente quatro textos paralelos entre as três línguas, totalizando doze ficheiros. Uma vez que se verificou que a quantidade de informação extraída destes ficheiros era suficiente para comprovar os resultados deste estudo, não havendo por isso, necessidade de sobrecarga de trabalho na validação dos resultados aplicados a um número superior de ficheiros.
Os textos seleccionados têm o formato de artigos legais, o que permite garantir que cada artigo numa língua tem a correspondente tradução nos respectivos ficheiros paralelos das outras línguas. Mas esta correspondência não é uma correspondência simples palavra a palavra, nem muitas vezes expressão mais complexa a expressão mais complexa, com manutenção da ordem dessas expressões.

A título ilustrativo, mostro a seguinte alínea de um artigo que ocorre num dos textos das três línguas estudadas.

Exemplo de texto em PT:

"**Quando bens ou serviços, estritamente necessários para o exercício das actividades oficiais da Organização ITER, forem adquiridos ou utilizados pela Organização ITER**, ou em seu nome, e quando o preço desses bens ou serviços inclua impostos ou direitos, a Parte toma, na medida do possível, as medidas adequadas para a concessão da isenção desses impostos ou direitos ou para a garantia do seu reembolso."

Exemplo de texto em ES:

"**Cuando se adquieran o utilicen en nombre de la Organización ITER bienes o servicios estrictamente necesarios para el ejercicio de sus actividades oficiales** y cuando el precio de estos bienes o servicios incluya impuestos o derechos, la parte correspondiente tomará, siempre que sea posible, las medidas adecuadas para conceder una exención de estos impuestos o derechos o para su reembolso."

Exemplo de texto em EN:

"**When goods or services, strictly necessary for the exercise of the official activities of the ITER Organization, are purchased or used by or on behalf of the ITER Organization**, and when the price of such goods or services includes taxes or duties, appropriate measures shall, whenever possible, be taken by the Party to grant exemption from such taxes or duties or to provide for their reimbursement."

Ao efectuarmos o alinhamento do início dos três textos (marcado a negrito) apercebemo-nos que os dois textos PT e EN estão na sua maioria alinhados sendo quase uma tradução directa um do outro, enquanto que o texto ES não representa uma tradução tão fiel. Este facto pode ser ilustrado na tabela seguinte, que representa o alinhamento dos três textos, onde estão marcadas a cinzento as zonas do texto desalinhadas.

EN	PT	ES
When	Quando	Cuando
		se adquieran o utilicen en nombre de la Organización ITER
Goods	bens	bienes
Or	ou	o
services,	serviços,	servicios
Strictly	estritamente	estrictamente
Necessary	necessários	necesarios
For	para	para
The	o	el
Exercise	exercício	ejercicio
Of	de	de
		sus
The	as	
Official		
Activities	actividades	actividades
	oficiais	oficiales
Of	de	
The	a	
ITER		
Organization	Organização	
	ITER	
,	,	
Are	forem	
Purchased	adquiridos	
Or	ou	
Used	utilizados	
By	por	
or on behalf of the ITER Organization		
	a Organização ITER, ou em seu nome	

Tabela 5: Alinhamento de três excertos de texto paralelos

Como foi dito anteriormente, para a obtenção de resultados foram utilizados quatro ficheiros para cada língua. A Tabela 6 e a Tabela 7 ilustram as dimensões do corpus usado (em número de palavras). O algoritmo para obtenção das traduções dos termos foi primeiramente aplicado aos textos de menores dimensões, obtendo-se os respectivos resultados. Posteriormente, aumentou-se o corpus em cerca de 60% [Tabela 7] ao processar o ficheiro de maiores dimensões (22006A1216_05), o que permitiu observar o comportamento das medidas de semelhança e do algoritmo perante um incremento significativo do número de termos a tratar.

Nome do Ficheiro	PT	EN	ES
22006A1216_05	3324	2914	3173
32005D0754	549	477	544
32006D0527	648	549	632
32006D0943	828	745	860

Tabela 6: Dimensão total dos ficheiros em termos por língua

Nome do Ficheiro	PT	EN	ES
22006A1216_05	62%	62%	61%
32005D0754	10%	10%	10%
32006D0527	12%	12%	12%
32006D0943	15%	16%	17%

Tabela 7: Distribuição percentual do total de termos dos ficheiros por língua

A Tabela 8 e a Tabela 9 mostram respectivamente a quantidade e percentagem de termos distintos em cada ficheiro por língua. Estes termos ocorrem somente num dos ficheiros e não nos outros, permitindo-nos quantificar o contributo de cada ficheiro para o aumento do corpus. Analisando a tabela de percentagens (Tabela 9), verifica-se para cada língua, que cerca de 80% do total dos termos constituintes do ficheiro de maiores dimensões (22006A1216_05) são repetidos nos outros ficheiros. Mas dado o elevado número de termos em questão, os cerca de 20% de termos distintos deste ficheiro, que são 685, 628 ou 719, respectivamente, para as línguas PT, EN e ES, são cerca de duas vezes e meia mais do que os 256, 240 e 252 termos distintos dos ficheiros "32006D0527" das línguas PT, EN e ES, que representam cerca de 40% de todos os temos destes ficheiros "32006D0527". Estes números demonstram que o contributo do ficheiro "22006A1216_05" para a construção do léxico é muito significativo, e tanto mais

significativo quando pensamos em função do número de termos completamente distintos ocorrendo nesses ficheiros com se mostra na Tabela 10 e na Tabela 11.

Nome do Ficheiro	PT	EN	ES
22006A1216_05	685	628	719
32005D0754	195	202	218
32006D0527	256	240	252
32006D0943	226	218	249

Tabela 8: Quantidade de termos distintos em cada ficheiro por língua

Nome do Ficheiro	PT	EN	ES
22006A1216_05	21%	22%	23%
32005D0754	36%	42%	40%
32006D0527	40%	44%	40%
32006D0943	27%	29%	29%

Tabela 9: Percentagem de termos distintos em cada ficheiro por língua (relativo ao número de termos)

As duas tabelas seguintes ilustram o contributo de cada ficheiro para o incremento dos termos constituintes do léxico, indicando a quantidade e percentagem de termos exclusivos de cada ficheiro por língua. Tomando como exemplo a versão portuguesa do ficheiro "22006A1216_05", isto significa que 71% dos termos que ocorrem no referido texto, não ocorrem nos demais ficheiros da mesma língua.

Nome do Ficheiro	PT	EN	ES
22006A1216_05	488	433	505
32005D0754	111	116	130
32006D0527	136	118	133
32006D0943	72	59	74

Tabela 10: Quantidade de termos exclusivos de cada ficheiro por língua

Nome do Ficheiro	PT	EN	ES
22006A1216_05	71%	69%	70%
32005D0754	57%	57%	60%
32006D0527	53%	49%	53%
32006D0943	32%	27%	30%

Tabela 11: Percentagem de termos exclusivos de cada ficheiro por língua (relativo ao número de termos)

3.2 Algoritmo Implementado

O algoritmo implementado é composto por duas fases, sendo a primeira de extracção dos termos, emparelhamento e classificação dos pares de traduções usando as medidas de *Dice* e *Levenshtein* [ver secção 2.3.9]. Após o processamento dos ficheiros, efectuam-se as respectivas validações das traduções e inserção no léxico dos pares novos. Numa segunda fase ocorre a extracção de novos termos por pivotagem incluindo as multi-palavras, seguindo-se igualmente o processo da validação das extracções feitas e da inserção dos termos válidos no léxico correspondente.

1) Efectuar o processamento de dois textos paralelos, separando ambos os textos em termos isolados. Neste passo, os sinais de pontuação são isolados das palavras a que estão ligados, dado que para este algoritmo não são usados como âncoras no alinhamento de termos nem sequer nas traduções.

2) Para os termos em Português (PT) e Espanhol (ES), decompor as contracções para normalizar os textos, dado que a contracção deixa a preposição que normalmente é gerida por alguma expressão à esquerda, enquanto o artigo depende da expressão da direita. Como exemplo temos a frase "o exercício **das** actividades oficiais" que após a decomposição da contracção "**das**" fica "o exercício **de as** actividades oficiais".

   Como as contracções PT são inúmeras, indico aqui apenas alguns exemplos de decomposição de contracções PT:

   do => de o
   duma => de uma
   nuns => em uns

Já a língua ES contém apenas duas contracções:

del => de el
al => a el

3) Inserir os termos de ambas as línguas numa base de dados, guardando também o offset relativo à distância do termo ao início do texto.

4) Criar todas as combinações[10] entre os termos de cada ficheiro a tratar e inserir na base de dados somente aquelas cujo valor da medida de *Dice* for superior a determinado limiar. Foi escolhido 0.5 como limiar o que significa que dois termos T_1 e T_2, com frequências de ocorrência $f(T_1)$ e $f(T_2)$, respectivamente pertencentes a um texto de uma língua L_1 e o correspondente texto paralelo escrito numa língua L_2, são aceites para inserção na base de dados, se a seguinte expressão for verdadeira:

$$dice(T_1, T_2) = 2 \cdot \frac{Min(f(T_1), f(T_2))}{f(T_1) + f(T_2)} \geq 0.5$$

Esta medida é uma adaptação da medida de *Dice* porque trabalho com o texto todo inicialmente.

5) Actualizar cada entrada inserida no passo anterior, com o valor da distância de *Levenshtein Normalizada* segundo a fórmula:

$$comp(T_1, T_2) = 1 - \frac{Lev(T_1, t_2)}{Max(Tamanho(T_1), Tamanho(T_2))}$$

Desta forma, cada par de termos (T_1, T_2), é classificado separadamente segundo as medidas de *Dice* [ver secção 2.3.4] e de *Levenshtein Normalizado* [ver secção 2.3.2].

6) Neste ponto resta efectuar a validação dos pares, classificando-os como válidos ou inválidos. Como foram geradas todas as combinações de termos, efectuou-se aqui um

[10] O espaço de pesquisa foi reduzido pela filtragem de posição relativa dos pares de termos, considerando somente os pares que não distam mais de 100 caracteres entre eles (valor empírico). Antes de executar o algoritmo aproveitou-se a informação recolhida nos ficheiros anteriores para validar as traduções já conhecidas, sempre considerando o referido filtro de posição. As traduções já validadas como certas ou erradas não são consideradas no processamento.

filtro de posição que rejeita os pares cuja distância é superior a 100 caracteres (este valor é empírico).
Nos casos em que para um mesmo termo existem vários candidatos, escolhe-se o par que apresente menor diferença entre posições relativas, significando que muito provavelmente estão relacionados entre si com traduções um do outro.
Estas associações de termos necessitam sempre de confronto com os textos respectivos aquando da validação.

7) Usando os pares validados como correctos ou incorrectos, aumenta-se o léxico com os novos termos.

A próxima fase irá extrair as traduções dos termos por pivotagem, permitindo enriquecer o léxico com termos que, de outro modo, não seriam detectados pelas medidas de semelhança. Recebe como entrada, os três textos paralelos de cada língua, os pares de termos validados na fase anterior e os respectivos léxicos.
Para exemplificar alguns dos passos do algoritmo considerou-se as línguas Português (PT), Espanhol (ES) e Inglês (EN), sendo PT a língua pivô e o par de termos a obter será (EN, ES).

Para actualizar o léxico com as entradas de pares (EN, ES) obtidos por pivotagem executa-se o seguinte algoritmo:

1) Seleccionar os pares ES-PT de traduções de termos validados (marcados como certos ou errados). No exemplo em baixo, devemos notar que este par de termos não foi considerado na primeira fase do algoritmo porque o valor da medida de *Levenshtein*Dice* é inferior a 0,5, desta forma, só será classificado usando pivotagem.

Termo PT	**Termo ES**	**Levenshtein**	**Dice**	**Levenshtein*Dice**
avaliações	evaluaciones	0,5	0,7692	0,3846

2) Seleccionar no léxico as entradas EN-PT de traduções cujos termos PT (“avaliações”) ocorrem nos pares ES-PT anteriores.

Termo EN	**Termo PT**
appraisal	avaliações
appraisals	avaliações
evaluations	avaliações

3) Usar somente as entradas de termos EN-PT cujo termo EN ocorra no respectivo texto dentro do intervalo de distância do termo PT, de forma a garantir que as traduções obtidas por pivotagem se restringem aos termos dos ficheiros em causa, evitando falsas traduções (traduções de termos correctos mas que não ocorrem nos textos em processamento).

Texto EN	Texto PT
2006/527/EC: Commission Decision of 27 July 2006 concerning the financing of studies and **evaluations** covering the areas of food safety, animal health and welfare and zootechnics.	2006/527/CE: Decisão de a Comissão de 27 de Julho de 2006 relativa a o financiamento de estudos e **avaliações** abrangendo as áreas de a segurança alimentar, saúde e bem-estar animal e zootecnia.

4) Como o termo inglês que existe simultaneamente no texto analisado e na listagem da alínea 2) corresponde à palavra "*evaluations*", podemos concluir que o termo "*evaluaciones*" em espanhol se traduz por "*evaluations*" em inglês pela relação que ambos partilham com o termo português "avaliações".

Na detecção de multi-palavras o algoritmo usa uma janela deslizante que tem um tamanho mínimo de dois termos (pela definição de multi-palavra).
A título de exemplo será utilizado o seguinte texto para ilustrar o funcionamento do algoritmo na detecção da tradução da multi-palavra espanhola "*seguridad alimentaria*" por "*food safety*" em inglês.

Texto EN	Texto PT	Texto ES
2006/527/EC: Commission Decision of 27 July 2006 concerning the financing of studies and evaluations covering the areas of **food safety**, animal health and welfare and zootechnics.	2006/527/CE: Decisão de a Comissão de 27 de Julho de 2006 relativa a o financiamento de estudos e avaliações abrangendo as áreas de a **segurança alimentar**, saúde e bem-estar animal e zootecnia.	2006/527/CE: Decisión de la Comisión de 27 de julio de 2006 relativa a la financiación de estudios e evaluaciones en los ámbitos de la **seguridad alimentaria**, la sanidad y el bienestar animal y la zootecnia.

Uma vez que seria muito moroso exemplificar a totalidade do algoritmo com este texto completo, o algoritmo irá ser demonstrado com o termo "de a segurança alimentar". Em

cada passo é mostrado o estado da janela deslizante, onde a coluna número de iteração, ajuda a que possamos identificar qual a janela deslizante corrente e a coluna de resultados indica qual foi o resultado do passo do algoritmo na iteração correspondente.

Para actualizar o léxico com multi-palavras de pares (EN, ES) obtidos por pivotagem, executa-se o seguinte algoritmo:

1) Percorrer o texto PT termo a termo da esquerda para direita, começando nos dois primeiros termos e avançando um termo a cada iteração. Os termos são inseridos na janela deslizante, pois serão usados para pesquisar as multi-palavras que contêm os termos da janela pela ordem com que foram inseridos.

Iteração Nº	**Janela Deslizante**		**Resultados (EN, PT)**
1	de	a	[estado inicial da janela]

2) Pesquisar no léxico os pares EN-PT onde a multi-palavra PT (na janela deslizante corrente) ocorra isoladamente ou ocorra inserida numa multi-palavra maior. Como o objectivo é encontrar a maior multi-palavra possível, a dimensão da janela deslizante não tem limite.

Iteração Nº	**Janela Deslizante**				**Resultados (EN, PT)**
1	de	a			Encontrou muitos resultados ir para (6)
2	de	a	segurança		Encontrou muitos resultados ir para (6)
3	de	a	segurança	alimentar	Não encontrou resultados ir para (3)
4	a	segurança	alimentar		Não encontrou resultados ir para (3)
5	segurança	alimentar			Encontrou um único resultado ir para (4)

3) Se nenhum par for encontrado, avança-se um termo e recomeça-se de (2).
NOTA: A janela deslizante tem sempre de ter no mínimo dois termos

4) Se for encontrado e for resultado único, verificar se este resultado é constituído exactamente pelos termos que se está a pesquisar e ir para (7).

Iteração Nº	Janela Deslizante		Resultados (EN, PT)
5	segurança	alimentar	Par ("***food safety***", "**segurança alimentar**") encontrado ir para (7)

5) Se o resultado não for o termo exacto, avança-se um termo e recomeça-se a partir de (2).

6) Se o resultado não é único, adiciona-se à pesquisa o próximo termo do texto e volta-se a (2). NOTA: Este passo é necessário para desambiguar os casos em que se tem mais de uma tradução para a mesma palavra.

7) Neste ponto só resta encontrar a expressão ES homóloga da expressão PT. Para isso são utilizados os pares de traduções ES-PT do texto, cujos termos PT ocorrem na multi-palavra PT. Desta forma, pode delimitar-se a multi-palavra ES pelos índices das palavras ES que ocorrem nas traduções ES-PT.

8) Partindo do par EN-PT de multi-palavras, foram identificados os pares de traduções de termos ES-PT cujos termos PT ocorrem na multi-palavra PT do par EN-PT. Com estes pares (ES-PT), e porque se sabe as posições dos termos no texto, é possível delimitar os termos ES que provavelmente (pode não existir tradução em ES para todos os termos PT) constituem a multi-palavra ES que será introduzida no léxico de pares EN-ES.

Multi-palavra PT	Termo PT	Termo ES
segurança alimentar	segurança	seguridad
	alimentar	alimentaria

Como os termos espanhóis "*seguridad*" e "*alimentaria*" se traduzem pelos termos portugueses "segurança" e "alimentar" respectivamente, e sabemos que a multi-palavra portuguesa "segurança alimentar" é traduzida por "*food safety*" em inglês, podemos inferir que a multi-palavra espanhola "*seguridad alimentaria*" que ocorre no texto é traduzida em inglês pela expressão "*food safety*".

9) Avançar um termo e recomeçar no ponto (2) até serem percorridos todos os termos do ficheiro.

Finalizando a identificação das multi-palavras, resta validar as novas entradas obtidas por pivotagem.

3.3 Precisão das Medidas de Semelhança

A presente secção é constituída por um conjunto de gráficos que descrevem os resultados das traduções extraídas pelo algoritmo apresentado nesta tese e ilustram a precisão das três medidas (*Levenshtein*, *Dice* e *Levenshtein*Dice*).
Mas, antes disso, numa primeira sequência de gráficos, mostra-se para cada par de línguas os resultados das validações das traduções geradas, sendo apresentados dois gráficos: um do número de traduções correctas e outro do número de traduções incorrectas. Estes gráficos representam a quantidade de termos traduzidos correctamente ou incorrectamente relativamente ao limiar de aceitação das traduções para determinado par de línguas, para cada uma das medidas de semelhança utilizadas. Como se pretende também analisar o comportamento do algoritmo e das medidas face ao aumento do corpus processado, são apresentados os gráficos com valores do processamento dos três ficheiros mais pequenos e posteriormente os mesmos gráficos incrementados com os resultados do tratamento do ficheiro de maiores dimensões.

Ao analisar as traduções do par PT-ES, os gráficos da Figura 12 e da Figura 13 representam a distribuição da quantidade de traduções em função do limiar de aceitação, as quais foram identificadas pelas medidas de semelhança e foram manualmente validadas, respectivamente como correctas ou incorrectas. Neste par de línguas, para limiares superiores a 0,9 todas as três medidas apresentam uma melhoria acentuada na detecção de traduções correctas, sendo a medida de *Dice* a melhor de todas.
Observando em simultâneo o gráfico das traduções incorrectas (Figura 13), verificamos que a medida de *Dice* apresenta também os maiores valores. Analisando somente o limiar 1,0, conclui-se que *Dice* detecta cerca de 100 traduções correctas a mais que as restantes medidas, mas em contrapartida, cerca de 350 traduções foram detectadas e consideradas erradas, contrastando bastante com as quase 0 (zero) de *Levenshtein* e *Levenshtein*Dice*, deste modo pode-se afirmar que a precisão da medida de *Dice* aproxima-se dos 45% para o par de línguas PT-ES, neste limiar de avaliação [Tabela 12].
Considerando uma outra perspectiva sobre estes dados, usando a medida de *Dice* validaram-se cerca de 650 traduções, entre correctas e incorrectas, o que difere das 200 traduções correctas obtidas pelas outras duas medidas, que a este nível de limiar extraem quase zero traduções incorrectas. A razão subjacente a estes resultados deve-se ao facto de *Dice* considerar somente as frequências de ocorrência dos termos nos respectivos textos, e *Levenshtein* basear-se na semelhança entre os termos de ambas as línguas.

Considerando as línguas em questão (Português e Espanhol) muitos dos termos são semelhantes.

A utilização da medida de *Dice* em conjunto com a de *Levenshtein* permitiu detectar traduções de termos que nunca seriam detectadas pela medida de *Levenshtein* para limiares superiores a 0,7, evitando contudo os problemas associados à medida de *Dice*. O termo PT "bem-estar" e o termo ES "*bienestar*" têm um valor de *Levenshtein Normalizado* de 0,67 e *Dice* de 1.

A utilização simultânea de ambas as medidas para a detecção de traduções, permite atenuar as lacunas das duas medidas. Por um lado *Levenshtein* detecta somente os termos semelhantes e tem uma precisão elevada (próxima de 100% para limiares superiores a 0,8), por outro lado *Dice* tem uma precisão muito mais baixa mas detecta traduções de termos com grafias muito distintas, baseando-se somente nas frequências dos termos.

A interpretação do resto do gráfico permite confirmar a análise anterior. Para os limiares inferiores a 0,9 verifica-se que a medida *Levenshtein*Dice* obtém ligeiramente melhores resultados que a medida de *Levenshtein* isolada e, consequentemente, produz menos traduções erradas. Para o limiar de 0,8 a medida de *Levenshtein* extrai erradamente cerca do dobro das traduções erradas da medida *Levenshtein*Dice* e detecta ligeiramente menos traduções correctas.

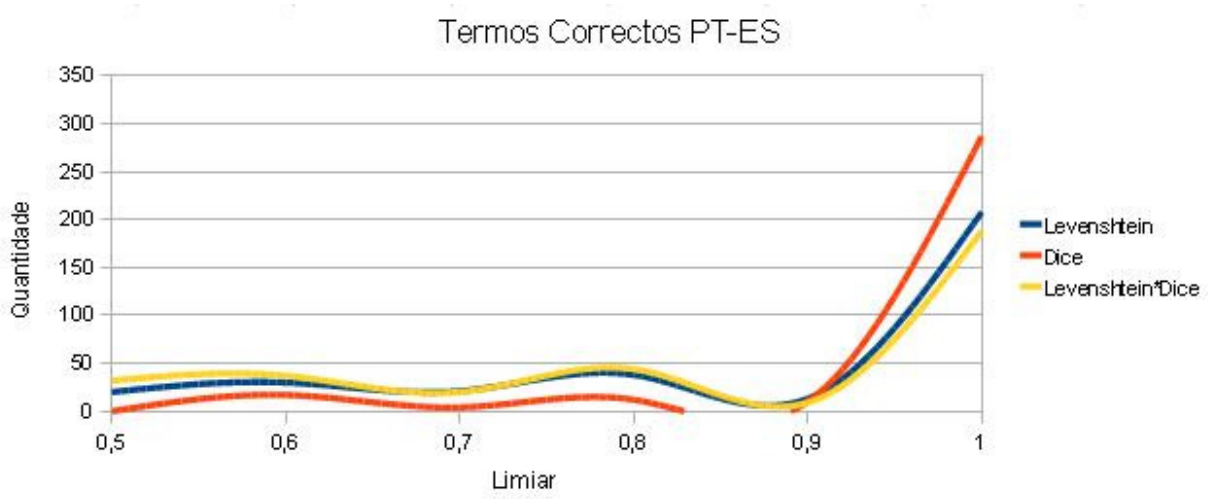

Figura 12: Termos correctos do par PT-ES antes de processar o ficheiro de maiores dimensões (22006A1216_05)

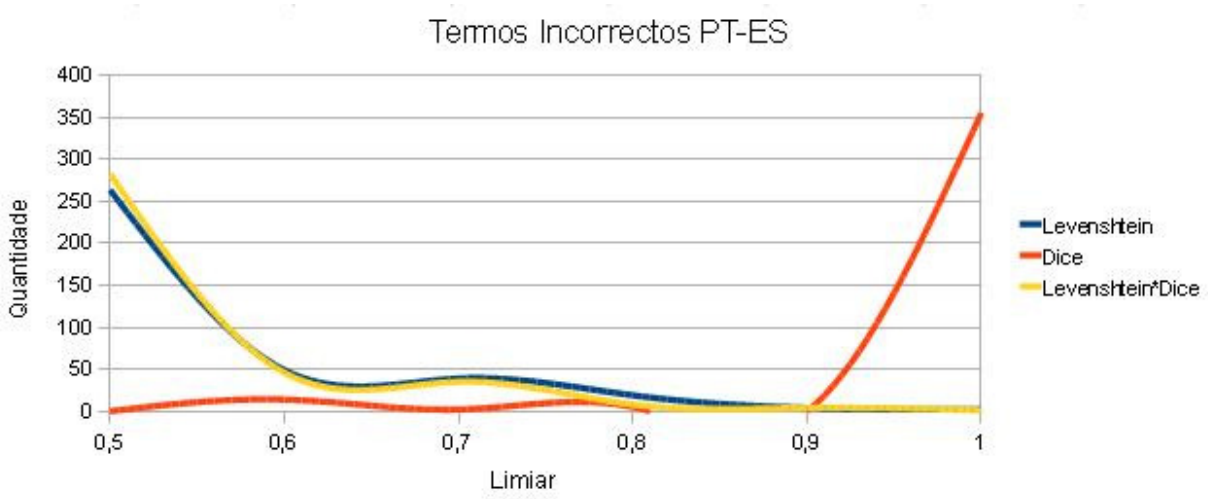

Figura 13: Termos incorrectos do par PT-ES antes de processar o ficheiro de maiores dimensões (22006A1216_05)

Como foi explicado na secção 3.1, o último texto a ser processado provocou um aumento de 60% do corpus. Continuando no par PT-ES, foram gerados os mesmos gráficos usando os resultados das traduções de todos os quatro ficheiros.

A característica mais evidente recai na semelhança dos gráficos, que mantêm as proporções dos anteriores. O facto de se trabalhar num corpus muito maior acentua alguns dos comportamentos das medidas e dos seus resultados discutidos anteriormente.

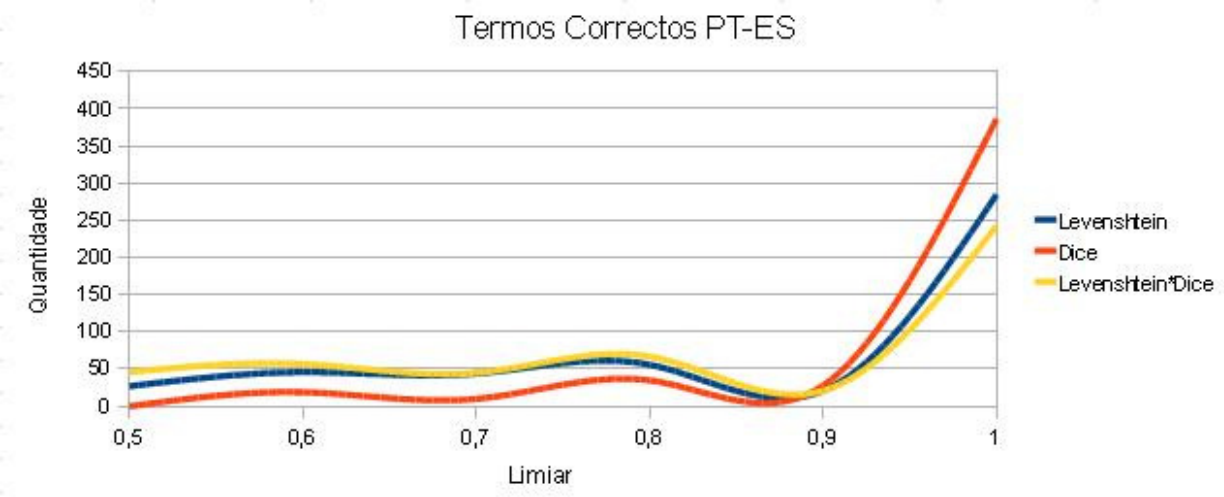

Figura 14: Termos correctos do par PT-ES depois de processar o ficheiro de maiores dimensões (22006A1216_05)

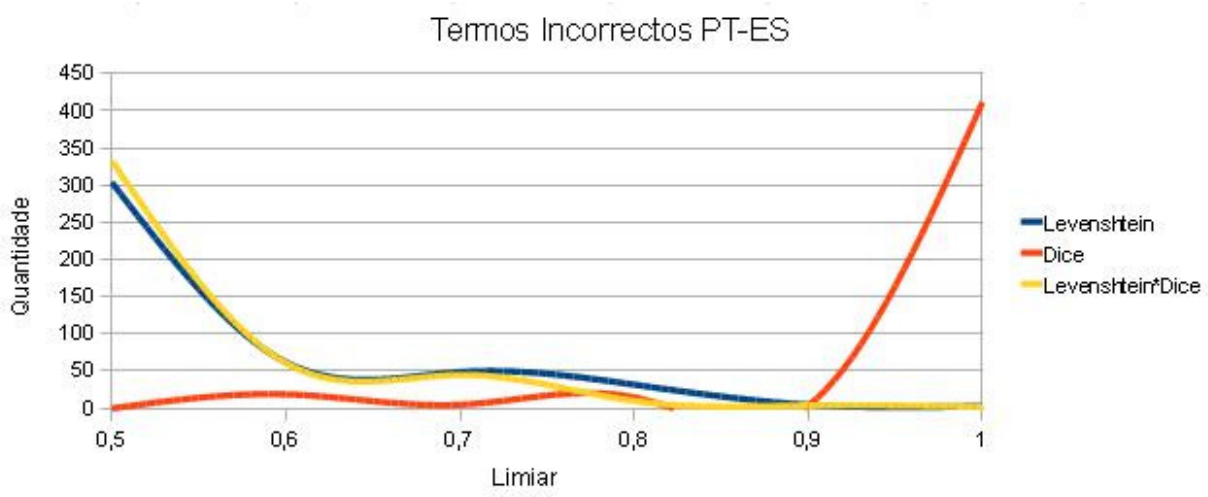

Figura 15: Termos incorrectos do par PT-ES depois de processar o ficheiro de maiores dimensões (22006A1216_05)

A semelhança entre as línguas Portuguesa e Inglesa é inferior à semelhança que existe entre Português e Espanhol. Esta característica dita os valores observados nos gráficos da Figura 16 e Figura 17. Comparativamente aos gráficos homólogos do par PT-ES, os valores das traduções correctas mantêm-se sensivelmente os mesmos para a medida de *Dice*, verificando-se a principal diferença a nível dos valores incorrectos que são somente cerca de 15% do total de termos validados, contra os mais de 50% do par PT-ES. *Levenshtein* mantém uma precisão de quase 100% para limiares superiores a 0,8. Para um limiar de 1,0, a medida de *Levenshtein* só detecta aproximadamente menos 50 traduções que o mesmo limiar no par PT-ES. Alguns destes termos são nomes próprios ou algarismos como as datas ou identificação dos artigos. Para limiares inferiores a 0,8 constata-se que o número de termos correctos é tendencialmente constante na orla dos 40 termos, apresentando valores de traduções incorrectas comparativamente mais baixos que para o par PT-ES, o que se deve à menor semelhança entre os pares PT-EN. Uma vez que os termos cognatos entre as duas línguas (PT-EN) são geralmente traduções correctas, entre PT-ES encontram-se alguns termos que diferem em género ou número ("profissionais/ *profesional*", "veterinários/*veterinaria*") devido à tradução feita pelo tradutor humano que escreveu o texto. Podem também ocorrer casos em que os dois termos têm um elevado grau de cognaticidade e significados diferentes, que por se encontrarem na mesma frase dão origem a traduções incorrectas, como nas frases exemplificativas em baixo, onde a palavra portuguesa "firma" que é sinónimo de empresa e a palavra espanhola "*firma*" que significa assinatura, são classificadas pela medida de *Levenshtein Normalizado* com o valor de 1.

Frase em português:

> " (…) as disposições do acordo serão aplicáveis a título provisório, nos termos da legislação interna da **firma** supracitada, aplicável na data da sua assinatura, enquanto se aguarda a sua entrada em vigor."

Frase em espanhol:

> " (…) las disposiciones del acuerdo se aplicarán de forma provisional con arreglo al derecho interno vigente en la empresa, a partir del día de su **firma**, a la espera de su entrada en vigor."

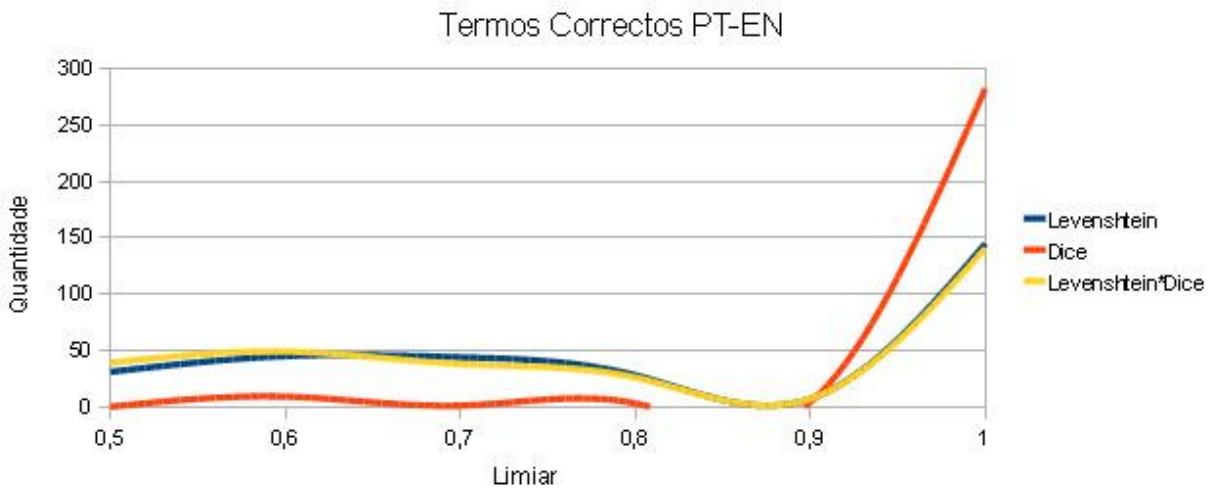

Figura 16: Termos correctos do par PT-EN antes de processar o ficheiro de maiores dimensões (22006A1216_05)

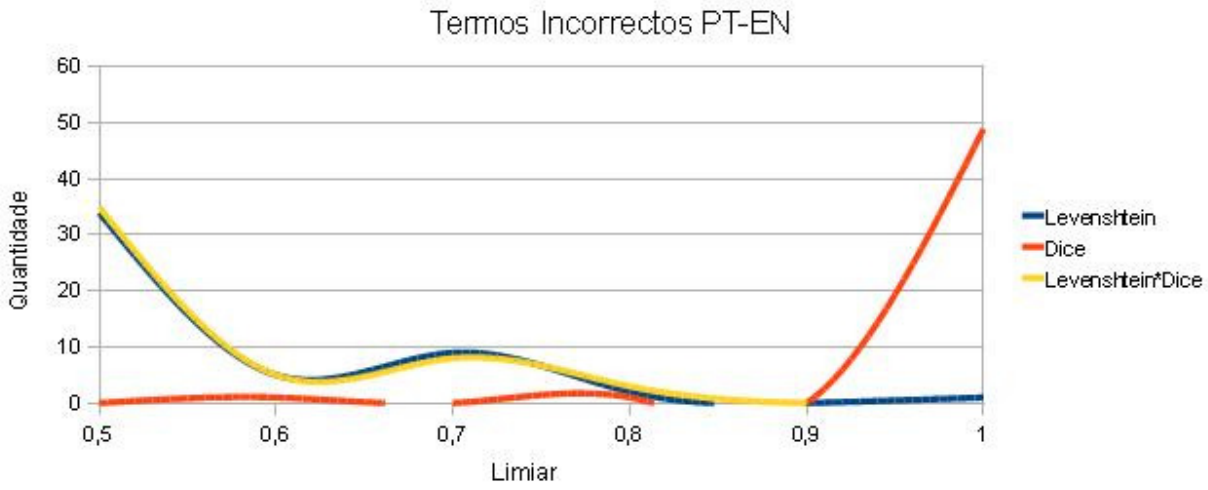

Figura 17: Termos incorrectos do par PT-EN antes de processar o ficheiro de maiores dimensões (22006A1216_05)

Relativamente ao processamento do ficheiro de maiores dimensões (Figura 18 e Figura 19), para estes pares de traduções PT-EN a única diferença a assinalar é o número de termos processados ser obviamente superior.
As traduções incorrectas mantêm os baixos valores identificados nos gráficos anteriores.

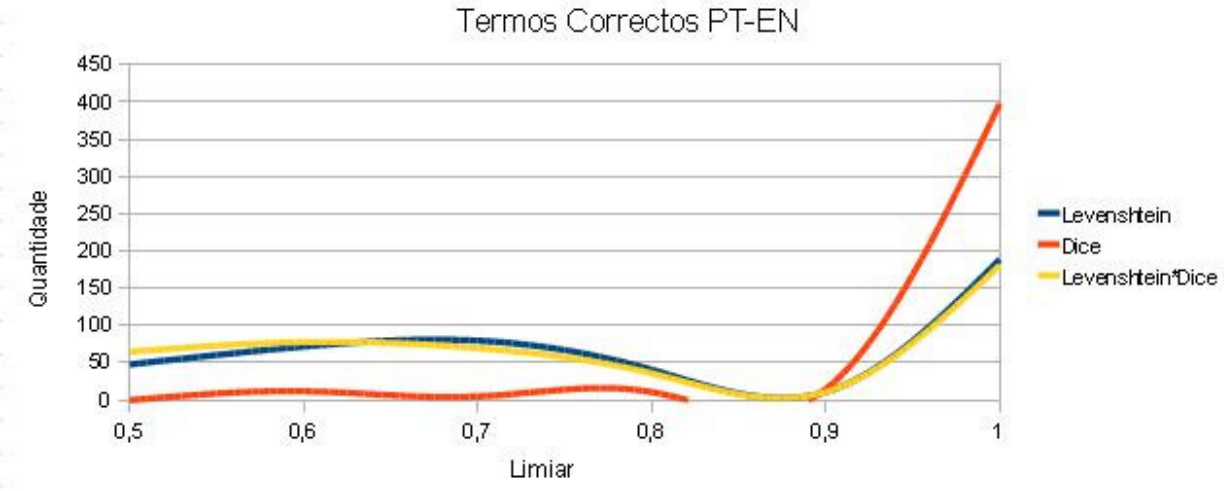

Figura 18: Termos correctos do par PT-EN depois de processar o ficheiro de maiores dimensões (22006A1216_05)

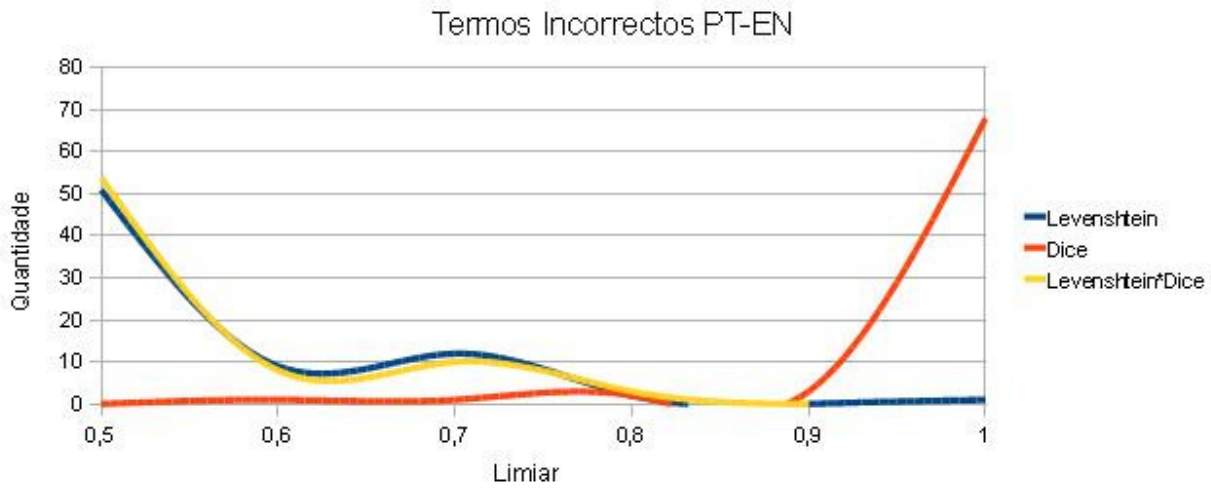

Figura 19: Termos incorrectos do par PT-EN depois de processar o ficheiro de maiores dimensões (22006A1216_05)

Analisando por último o par ES-EN (Figura 20 e Figura 21), continua-se a observar a mesma tendência dos gráficos do par PT-EN com excepção dos valores do par ES-EN serem globalmente mais baixos em cerca de metade. A análise efectuada ao par PT-EN aplica-se também a este par de línguas (ES-EN).

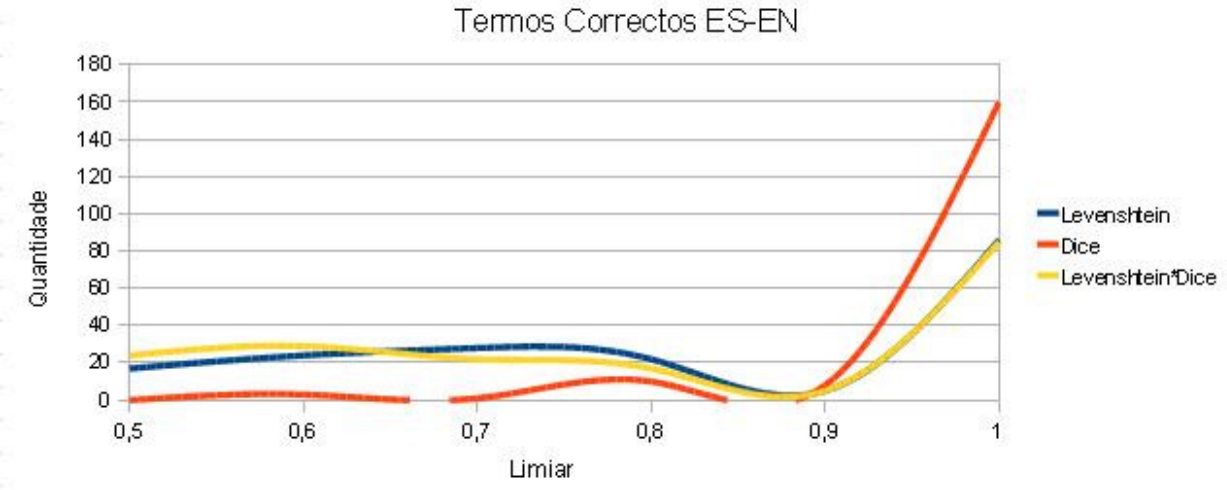

Figura 20: Termos correctos do par ES-EN antes de processar o ficheiro de maiores dimensões (22006A1216_05)

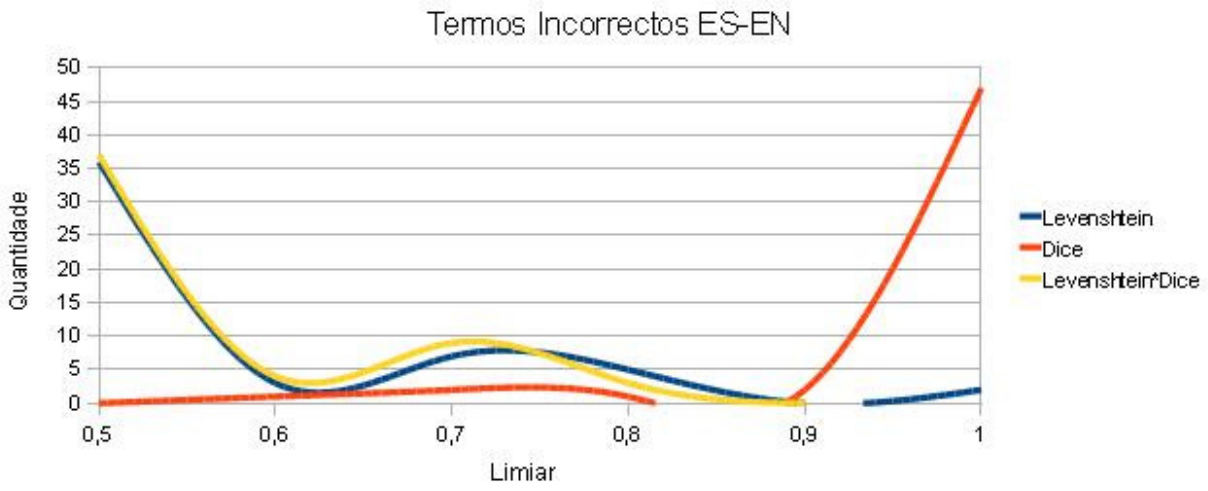

Figura 21: Termos incorrectos do par ES-EN antes de processar o ficheiro de maiores dimensões (22006A1216_05)

Os gráficos da Figura 22 e Figura 23 mantêm, como seria de esperar, as mesmas características que os demais gráficos criados a partir dos resultados do processamento dos quatro ficheiros, pelo que a análise será semelhante.

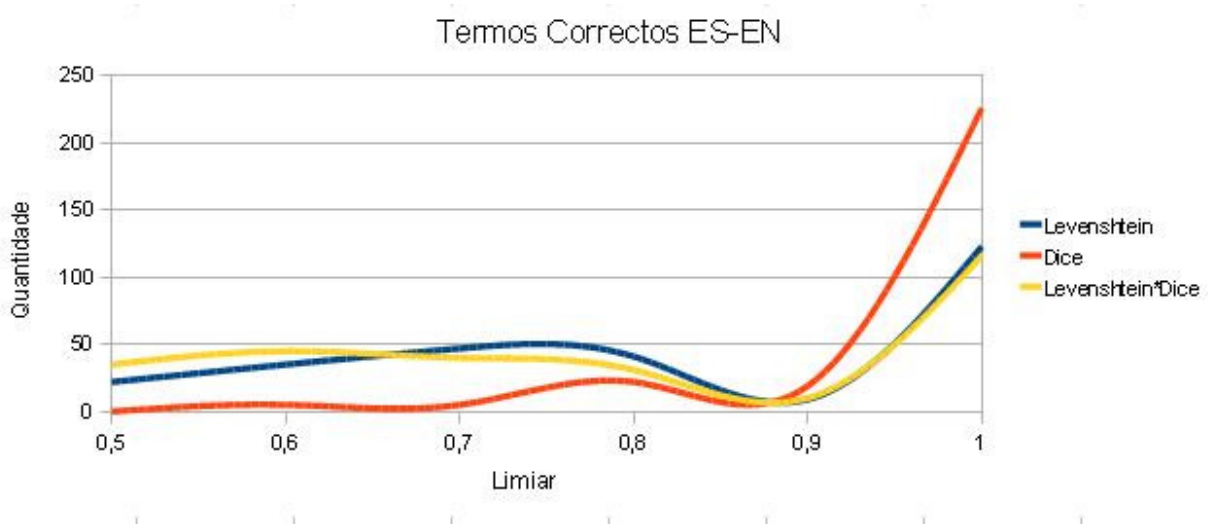

Figura 22: Termos correctos do par ES-EN depois de processar o ficheiro de maiores dimensões (22006A1216_05)

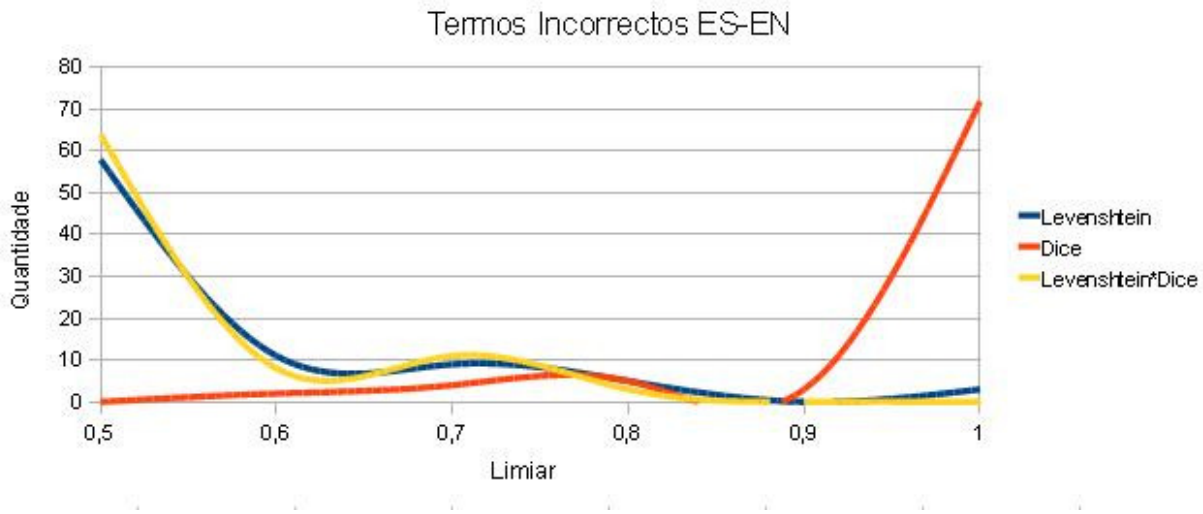

Figura 23: Termos incorrectos do par ES-EN depois de processar o ficheiro de maiores dimensões (22006A1216_05)

3.3.1 Precisão da Medida de *Dice*

Os gráficos anteriores permitiram analisar os resultados e comparar as medidas em função da quantidade de traduções avaliadas, mas a melhor forma de identificar a precisão das medidas de semelhança usadas, consiste na normalização dos resultados transformando-os em percentagens.

Os gráficos e tabelas seguintes representam a distribuição percentual das traduções correctas **em cada limiar** por cada um dos três pares de línguas estudados. Ou seja, os valores considerados para um limiar de *Dice* igual a 0,5 são os que estão contidos no intervalo [0,5; 0,6 [, sucedendo de forma semelhante para cada limiar de aceitação.

Para cada uma das três medidas usadas, são gerados dois gráficos de precisão. Um com os dados de precisão da medida analisada, recolhidos do processamento dos três ficheiros mais pequenos, e outro com os dados do total dos quatro ficheiros (incluindo o ficheiro de maiores dimensões) para que se possa avaliar o comportamento das medidas perante um aumento significativo do corpus.

Limiar	PT-EN	ES-EN	PT-ES
0,50	0,01%	0,01%	0,01%
0,60	0,03%	0,03%	0,03%
0,70	90,91%	57,14%	54,84%
0,80	100,00%	80,00%	61,54%
0,90	85,71%	87,50%	87,50%
1,00	85,24%	77,29%	44,98%

Tabela 12: Precisão da medida de Dice por par de línguas antes de processar o ficheiro de maiores dimensões (22006A1216_05)

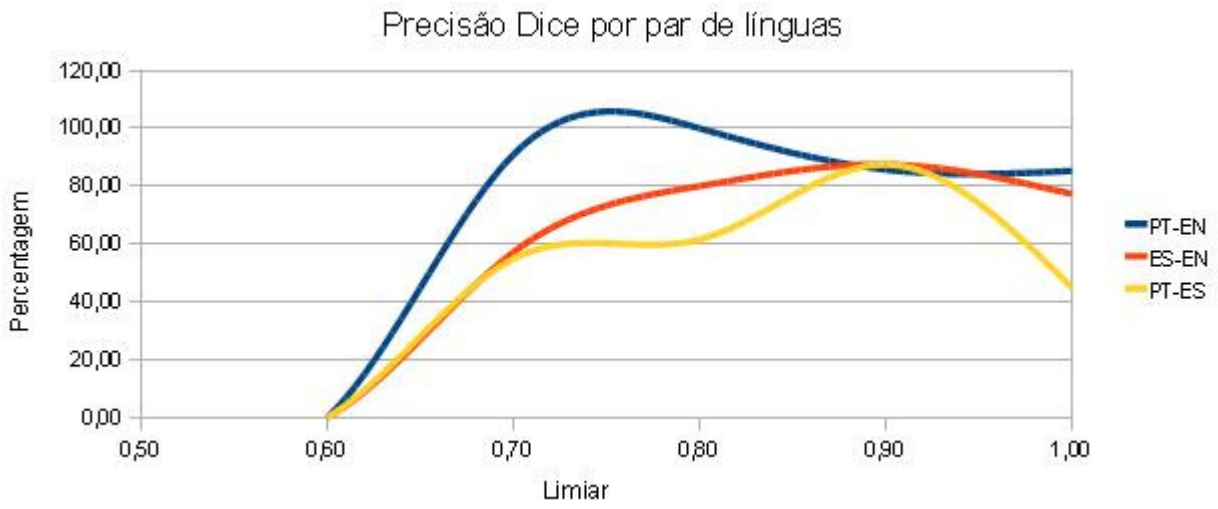

Figura 24: Precisão da medida de Dice por par de línguas antes de processar o ficheiro de maiores dimensões (22006A1216_05)

Limiar	**PT-EN**	**ES-EN**	**PT-ES**
0,50	0,01%	0,01%	0,01%
0,60	0,03%	50,00%	0,03%
0,70	93,75%	70,00%	51,16%
0,80	75,00%	64,71%	68,57%
0,90	85,00%	87,50%	75,00%
1,00	85,38%	76,16%	49,26%

Tabela 13: Precisão da medida de Dice por par de línguas depois de processar o ficheiro de maiores dimensões (22006A1216_05)

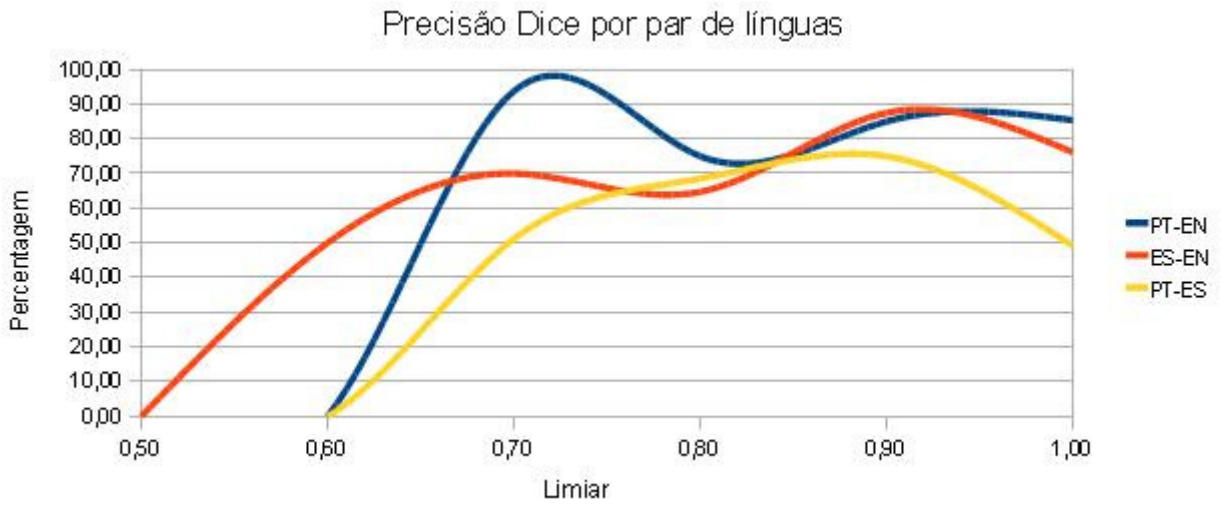

Figura 25: Precisão da medida de Dice por par de línguas depois de processar o ficheiro de maiores dimensões (22006A1216_05)

3.3.2 Precisão da medida de *Levenshtein*

Esta secção apresenta os resultados da precisão da medida *Levenshtein* e, à semelhança da secção anterior, regista os resultados da aplicação da medida aos termos dos textos processados, indicando a precisão por par de línguas, anterior e posterior ao processamento o ficheiro de maiores dimensões (22006A1216_05).

Limiar	PT-EN	ES-EN	PT-ES
0,50	33,33%	25,00%	5,69%
0,60	90,38%	78,79%	25,32%
0,70	87,80%	90,00%	41,18%
0,80	80,85%	72,22%	44,87%
0,90	95,00%	88,89%	74,42%
1,00	99,32%	97,73%	99,05%

Tabela 14: Precisão da medida de Levenshtein por par de línguas antes de processar o ficheiro de maiores dimensões (22006A1216_05)

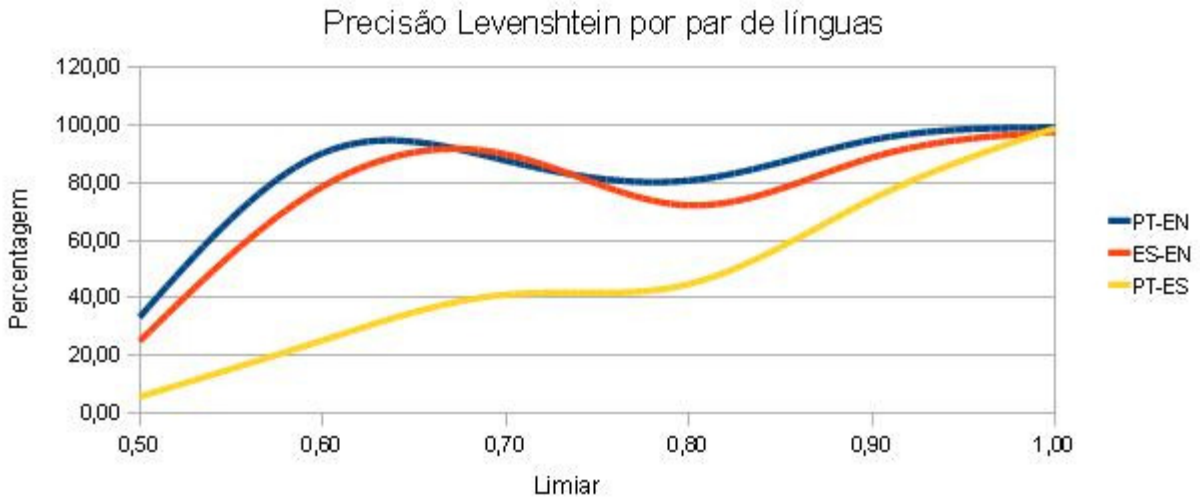

Figura 26: Precisão da medida de Levenshtein por par de línguas antes de processar o ficheiro de maiores dimensões (22006A1216_05)

Após o processamento do ficheiro de maiores dimensões (22006A1216_05) obtém-se os resultados mostrados na tabela e gráfico seguintes.

Limiar	PT-EN	ES-EN	PT-ES
0,50	36,23%	20,00%	6,12%
0,60	87,34%	68,75%	28,44%
0,70	89,19%	84,21%	46,67%
0,80	85,14%	80,33%	50,42%
0,90	96,67%	93,33%	78,46%
1,00	99,47%	97,62%	98,96%

Tabela 15: Precisão da medida de Levenshtein por par de línguas depois de processar o ficheiro de maiores dimensões (22006A1216_05)

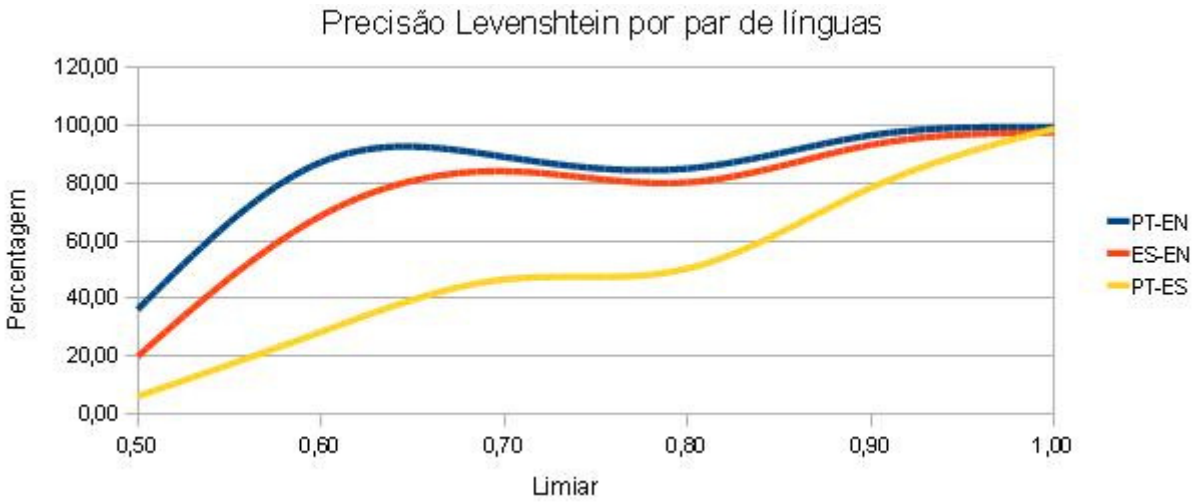

Figura 27: Precisão da medida de Levenshtein por par de línguas depois de processar o ficheiro de maiores dimensões (22006A1216_05)

3.3.3 Precisão da medida de *Levenshtein*Dice*

Esta secção apresenta os resultados da precisão da medida combinada *Levenshtein*Dice*.

Limiar	PT-EN	ES-EN	PT-ES
0,50	41,51%	29,55%	7,20%
0,60	90,57%	79,49%	33,33%
0,70	88,89%	84,62%	44,00%
0,80	79,49%	68,97%	54,29%
0,90	90,00%	91,67%	80,56%
1,00	100,00%	100,00%	99,48%

Tabela 16: Precisão da medida de Levenshtein*Dice por par de línguas antes de processar o ficheiro de maiores dimensões (22006A1216_05)

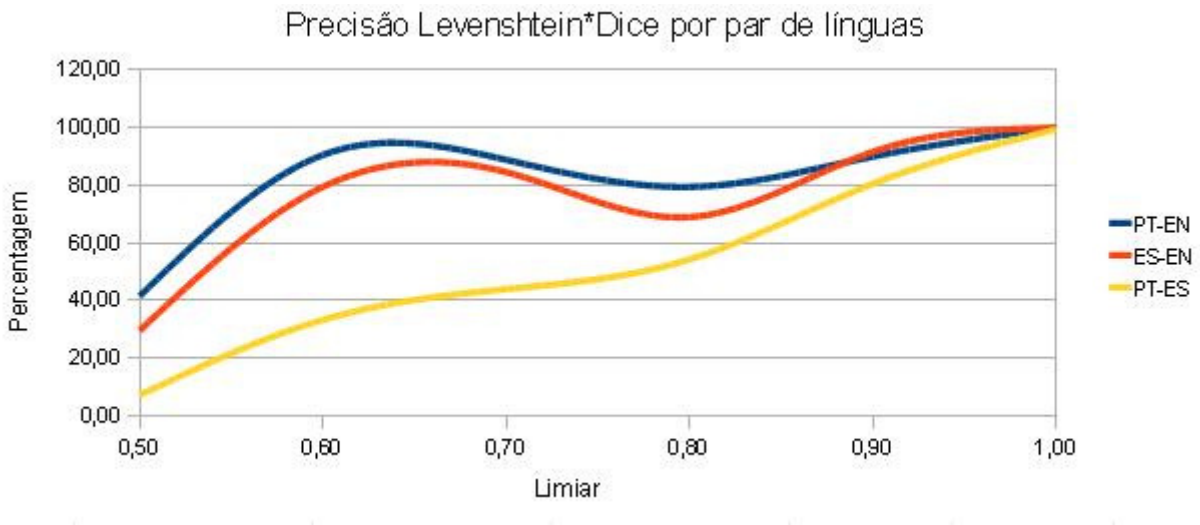

Figura 28: Precisão da medida de Levenshtein*Dice por par de línguas antes de processar o ficheiro de maiores dimensões (22006A1216_05)

Comparando a Tabela 16 com a Tabela 17, verifica-se que a partir de um limiar de 0,7 se nota uma melhoria de precisão (cerca de 3% a 4%) para a medida de *Levenshtein*Dice* resultante do aumento do tamanho do corpus.

Limiar	PT-EN	ES-EN	PT-ES
0,50	43,02%	23,94%	8,71%
0,60	91,36%	73,33%	35,82%
0,70	91,14%	90,91%	51,81%
0,80	83,05%	78,43%	61,46%
0,90	93,10%	95,00%	88,52%
1,00	100,00%	100,00%	99,20%

Tabela 17: Precisão da medida de Levenshtein*Dice por par de línguas depois de processar o ficheiro de maiores dimensões (22006A1216_05)

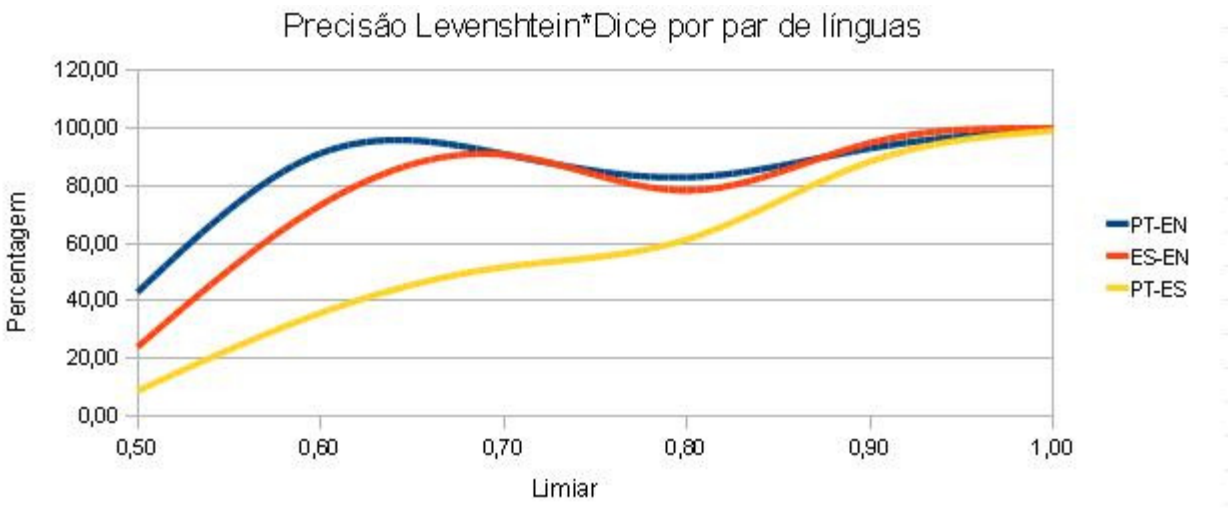

Figura 29: Precisão da medida de Levenshtein*Dice por par de línguas depois de processar o ficheiro de maiores dimensões (22006A1216_05)

3.3.4 Comparação das três medidas

As tabelas e gráficos apresentados na secção 3.3 permitem analisar as características e os comportamentos das medidas de semelhança estudadas, mas utilizam valores agregados por intervalo de limiar[11], tornando difícil sugerir correctamente o limiar a utilizar. Desta forma, para retirar conclusões sobre o limiar a eleger, existiu a necessidade de agregar os dados de outra maneira. Assim, a Tabela 18 e a Figura 30 disponibilizam uma visão acumulada dos resultados por medida, ou seja, ao limiar de 0,5 correspondem todos os valores cujo limiar é superior ou igual a 0,5.

Limiar	Levenshtein*Dice	Levenshtein	Dice
0,50	65,70 %	34,02 %	0,49 %
0,60	86,42 %	67,57 %	66,02 %
0,70	90,92 %	80,59 %	66,23 %
0,80	97,01 %	89,64 %	66,14 %
0,90	98,98 %	95,54 %	65,61 %
1,00	99,63 %	97,09 %	64,68 %

Tabela 18: Precisão das medidas de semelhança com valores acumulados

[11] Ex: ao limiar 0,5 correspondem os valores contidos no intervalo [0,5; 0,6[

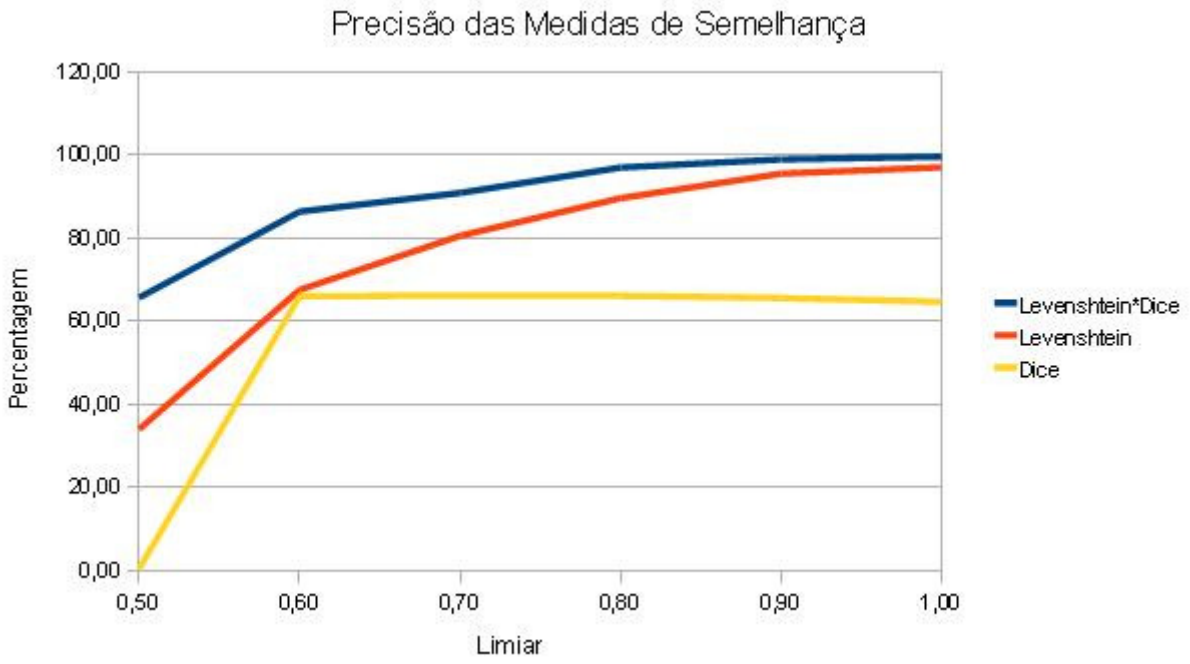

Figura 30: Precisão das medidas de semelhança com valores acumulados

O gráfico acima evidencia bem a excelência da medida combinada de *Levenshtein* e *Dice*, face à utilização isolada de cada uma. A medida de *Dice*, como seria de esperar, apresenta a pior precisão, tornando-se praticamente constante para limiares a partir de 0,6. Os resultados da medida de *Levenshtein* mostram o destaque da precisão desta medida em limiares superiores a 0,6 os quais ficam acima dos 80%. Este comportamento é justificado, pelo facto desta medida identificar palavras cognatas, e por se trabalhar só com línguas europeias, que partilham a origem de muitas das suas palavras.
Como a medida *Levenshtein*Dice* resulta da combinação das medidas *Levenshtein* e *Dice*, é espectável que absorva as melhores características de ambas[12], levando a sua precisão a destacar-se em todos os limiares.

Dado que este estudo incide sobre três pares de línguas distintos, convém mostrar os resultados agregados por par de línguas, para que se possa apreciar melhor os pormenores a eles associados. Desta forma, os gráficos e tabelas seguintes representam os valores obtidos para a medida de *Levenshtein* (Tabela 19 e Figura 31) e para a medida de *Levenshtein*Dice* (Tabela 20 e Figura 32), acumulados por limiar e agregados por par de línguas (ES-EN, PT-EN e PT-ES).

[12] Cognaticidade de *Levenshtein* e frequência das ocorrências dos termos de *Dice*.

Limiar	ES-EN	PT-EN	PT-ES
0,50	32,18	49,89	26,97
0,60	71,19	84,81	56,03
0,70	84,41	92,42	71,53
0,80	92,11	98,39	83,60
0,90	95,65	99,50	93,07
1,00	96,85	99,48	95,68

Tabela 19: Precisão de *Levenshtein* com valores acumulados agregados por par de línguas

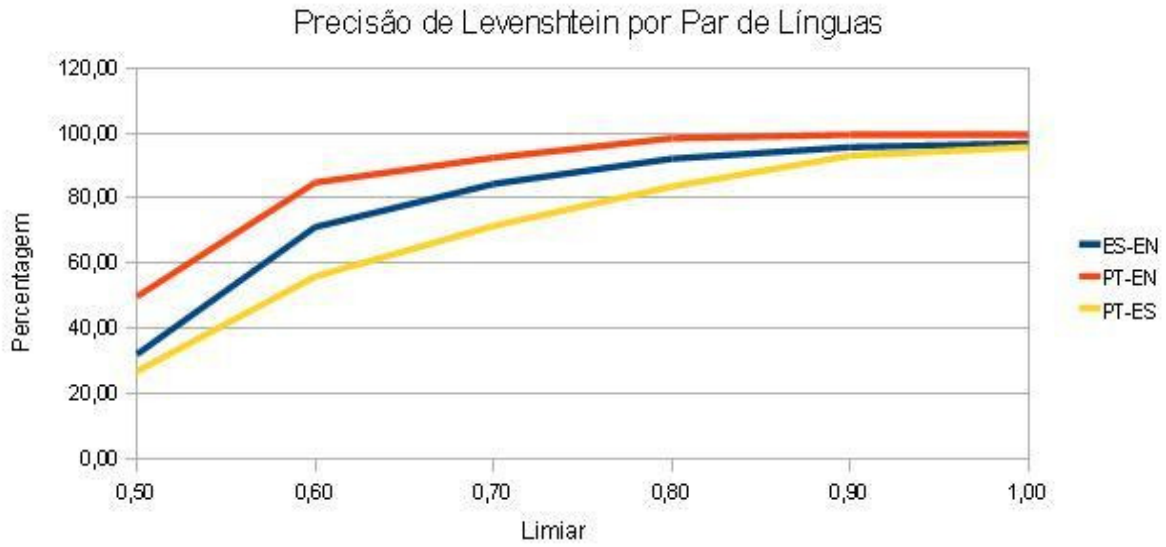

Figura 31: Precisão de *Levenshtein* com valores acumulados agregados por par de línguas

No gráfico anterior, ressaltam os resultados sobre o par PT-ES, que apresenta os piores resultados dos três pares de línguas. Esta situação deve-se ao facto de bastantes termos serem traduzidos por termos de número diferente (um termo que esteja plural numa língua ser traduzido por um termo da outra língua no singular) [ver secção 3.4]. O par PT-EN apresenta valores acima dos 90% para limiares superiores ou iguais a 0,7, sendo este o par que apresenta melhores resultados, motivados pela menor semelhança entre estas duas línguas, o que origina estes valores elevados de cognaticidade.

A medida adoptada para classificar os pares de termos é Levenshtein*Dice, pelo que não podemos deixar de mostrar os resultados desta medida, que são apresentados pela tabela e gráfico seguintes.

Limiar	ES-EN	PT-EN	PT-ES
0,50	76,10	85,49	50,90
0,60	91,63	94,71	78,02
0,70	93,81	95,82	86,04
0,80	98,11	98,70	95,39
0,90	100,00	100,00	97,77
1,00	100,00	100,00	99,18

Tabela 20: Precisão *Levenshtein*Dice* com valores acumulados por par de línguas

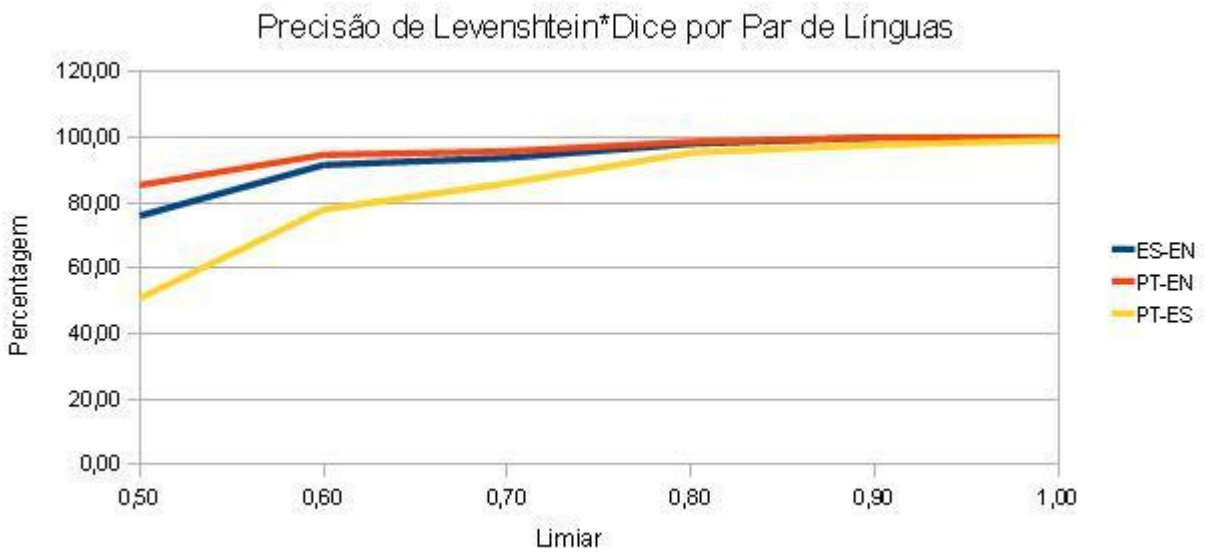

Figura 32: Precisão Levenshtein*Dice com valores acumulados por par de línguas

O comportamento do par PT-ES mantém-se como o resultado menos promissor, não invalidando que apresente também ele, muito bons resultados.
Comparativamente com *Levenshtein*, verifica-se um aumento extremamente significativo da precisão na detecção de traduções ao utilizarmos esta medida (*Levenshtein*Dice*). Para os pares ES-EN e PT-EN conseguem-se inclusivamente, precisões acima dos 90%

para limiares superiores ou iguais a 0,6. Com textos do par PT-ES os resultados obtidos surpreendem, porque não se esperavam valores de precisão tão baixos. Contudo, estes resultados percebem-se à luz da maior liberdade de tradução dos textos em espanhol.

Para finalizar esta secção e a título de conclusão, analisa-se a seguinte tabela das percentagens de traduções correctas usando a medida *Levenshtein*Dice* antes e depois de processar o ficheiro de maiores dimensões (22006A1216_05).

	PT-ES	PT-EN	ES-EN
Antes de 22006A1216_05	46.68%	85.47%	77.45%
Depois de 22006A1216_05	51.28%	85.47%	76.31%

Tabela 21: Percentagem de traduções correctas usando a medida Levenshtein*Dice

Dado que o Português e o Espanhol são duas línguas tão próximas, é espectável que haja uma melhoria na percentagem de traduções validadas como correctas. Este facto é visível na Tabela 21, que para o par PT-ES apresenta um incremento de cerca de 5% de traduções correctas.

3.4 Precisão da Pivotagem

O enriquecimento do léxico foi efectuado em várias etapas [ver secção 3.2], intercaladas por validações das traduções extraídas.
No decorrer desta tese foram extraídas traduções por dois métodos: usando as medidas de semelhança *Dice*, *Levenshtein* e *Levenshtein*Dice*, e usando pivotagem com três pares de línguas em simultâneo. Através das medidas de semelhança, extraíram-se traduções para os pares de línguas PT-EN, ES-EN e PT-ES. Como no início dos trabalhos somente se possuía o léxico PT-EN já com uma dimensão considerável, contendo traduções de termos e multi-termos deste par de línguas, a primeira extracção de traduções de termos para os pares restantes (ES-EN e PT-ES), permitiu após a respectiva validação, criar um léxico inicial para os pares ES-EN e PT-ES.
Dado que as medidas de semelhança usadas têm dificuldade em extrair termos com grafias distintas, e porque possui um léxico PT-EN com cerca de 180 000 entradas validadas, consegue-se extrair por pivotagem, termos isolados para o par ES-EN usando PT como língua pivô. Após esta fase procedeu-se também à validação das traduções obtidas.

Como o léxico PT-EN possui multi-termos, usando também a pivotagem obtiveram-se alguns multi-termos[13] para o par ES-EN, que de outro modo não seriam obtidos pelas técnicas usadas anteriormente.
Observando a Tabela 22, que ilustra a precisão obtida pela pivotagem, nota-se que cerca de 43% do total de entradas extraídas para o par ES-EN, foram traduzidas por pivotagem. O total de termos válidos extraídos para este par ronda os 81%, demonstrando claramente o contributo da pivotagem para estes resultados, especialmente se compararmos com o par PT-EN, para o qual não foi usada pivotagem e motivou os 64% de termos válidos.

Par de Línguas	**Estado dos Termos**		**Número de Entradas**	**% por Par de Línguas**
PT-EN	Termos Inválidos		74	35,75%
	Termos Válidos		133	64,25%
ES-EN	Termos Inválidos		68	18,89%
	Termos Válidos	Medidas de Semelhança	135	37,50%
		Pivotagem	157	43,61%
PT-ES	Termos Inválidos		442	58,01%
	Termos Válidos		320	41,99%

Tabela 22: Precisão da Pivotagem

Os valores para o par de línguas PT-ES são díspares dos resultados dos outros pares. Das traduções extraídas 58% foram classificadas como erradas e só 42% estão certas. Esta situação deve-se ao facto de bastantes termos serem traduzidos por termos de número diferente (um termo que esteja plural numa língua ser traduzido por um termo da outra língua no singular).
No exemplo apresentado a seguir, extraído de um dos textos em português pode verificar-se entre outros, o termo "específico" traduzido por "*específicos*" no texto espanhol, "sujeita" por "*sujetas*, "acordo" por "*acuerdos*", além de outras alterações observáveis.

[13] Dos termos válidos do par ES-EN obtidos por pivotagem, 32 são multi-termos.

Texto PT:
"(5) Cada uma das tarefas deve ser **sujeita** a um **acordo específico**. Estes acordos serão assinados entre a Comissão e o contratante seleccionado, tal como definido no contrato-quadro."

Texto ES:
"(5) Todas las tareas concretas están **sujetas** a **acuerdos específicos**, que deben ser firmados por la Comisión y el contratista seleccionado con arreglo a lo dispuesto en el contrato marco."

3.5 Comparação com Outros Autores

O trabalho desenvolvido por Charles Shafer e David Yarowsky [ver secção 2.2.5] apresenta um método de indução de léxicos entre duas línguas distantes, sem necessidade de quaisquer corpora paralelos bilingues. O algoritmo combina com a similaridade de ocorrência temporal entre as datas em corpora das notícias, similaridade do contexto entre línguas, a distância de *Levenshtein* ponderada, a frequência relativa e medidas de similaridade "*burstiness*", não necessitando um dicionário entre o inglês e a língua de destino, no entanto, necessita um dicionário de dimensão importante entre a língua pivô e o Inglês.
O facto de necessitar de um dicionário/léxico com dimensão considerável, entre a língua pivô e uma das outras línguas, é a única semelhança com o trabalho desenvolvido nesta dissertação, dado que o trabalho apresentado por Shafer utiliza uma combinação de medidas distintas, não sendo por isso comparável.

Gideon S. Mann e David Yarowsky [ver secção 2.2.4] apresentaram um método de indução de léxicos que permite relacionar cognatos de pares de línguas através de uma língua pivô. Os léxicos bilingues entre línguas da mesma família são induzidos utilizando modelos probabilísticos de distância entre cognatos de textos paralelos, em particular a distância de *Levenshtein*. Os léxicos para tradução entre pares de línguas de raízes distintas, foram gerados por uma combinação destes modelos de tradução intra-familiar, e um ou mais dicionários on-line para línguas com bases diferentes. Em média os dicionários continham inicialmente cerca de 900 entradas.
Obtiveram até 95% de precisão no vocabulário de destino, permitindo desta forma, que partes substanciais dos léxicos possam ser geradas com precisão, para idiomas que não possuam dicionários bilingues ou corpora paralelos.

Foi mostrado que línguas da mesma família são próximas o suficiente, de modo a que os pares de cognatos entre duas línguas são comuns, e porções significativas do léxico podem ser induzidas com alta precisão.

Os resultados obtidos nesta dissertação são comparáveis e muito semelhantes aos apresentados por Mann e Yarowsky, que obtiveram cerca de 92% de precisão nas extracções de traduções para o par PT-ES usando a medida de *Levenshtein*. Conseguiram-se nesta tese precisões de *Levenshtein* próximas de 90% para o par PT-ES (o único par que temos em comum) em limiares de aceitação superiores a 0,8 [Figura 31], mas para os restantes pares de línguas (PT-EN e ES-EN) obtiveram-se resultados muito superiores atingindo os 90% para limiares superiores a 0,7. No entanto, um dos contributos desta tese é a utilização da medida *Levenshtein*Dice*, onde se conseguiram precisões para os pares de línguas PT-EN e ES-EN, superiores a 90% para limiares superiores a 0,6 [Figura 32], enquanto que para o par PT-ES só se obtêm estes resultados em limiares superiores a 0,7. Ao introduzir a pivotagem, foi possível enriquecer o léxico para o par ES-EN usando PT como língua pivô. As entradas do léxico obtidas por pivotagem, são constituídas por palavras isoladas e por multi-palavras de qualidade (precisão perto de 100%). O léxico final extraído para este par de línguas é constituído por traduções de palavras extraídas usando a medida de semelhança e por palavras e multi-palavras extraídas por pivotagem, obtendo cerca de 81% de precisão final [ver Tabela 22]. Este valor fica um pouco aquém do resultado apontado por Mann e Yarowsky que é de 95% de precisão no vocabulário de destino, mas iniciaram a extracção com dicionários já existentes, facto que nesta dissertação não era premissa, uma vez que a extracção se iniciou somente com um léxico para um dos pares trabalhados, pelo que os restantes léxicos, em particular o léxico PT-ES, foram extraídos de raiz recorrendo às técnicas já apresentadas.

4. Conclusões e Trabalho Futuro

A construção automática de léxicos bilingues exige bastante esforço humano nas validações das traduções de termos extraídas. Um algoritmo que extraia pares de termos (propostas de traduções) com uma taxa elevada de erro, aumenta o esforço na validação das traduções. Assim sendo, um algoritmo para criação automática de léxicos bilingues, não só terá de extrair muitas entradas novas, como garantir uma elevada precisão nas traduções obtidas.

Esta dissertação apresenta um método de construção automática de léxicos bilingues com recurso a pivotagem com utilização simultânea de três pares de línguas, e três medidas de semelhança (*Levenshtein Modificado*, *Dice* e *Levenshtein*Dice*) para efectuar a detecção de traduções de termos. O algoritmo criado é muito simples e relativamente eficiente,

apresentando uma precisão da medida de semelhança *Levenshtein*Dice* superior a 90% para um limiar de aceitação de 0,7 e atingindo aproximadamente 100% para limiares próximos de 1,0 [ver Tabela 20]. A introdução da pivotagem foi motivada pelo facto de já existir um léxico validado para um dos pares de línguas (PT-EN), permitindo a extracção de termos e multi-termos que de outro modo não seriam obtidos. O facto da pivotagem ter sido executada com três pares de línguas em simultâneo, permitiu verificar que contribui para um aumento significativo da precisão, dado que cerca de 43,5% das entradas válidas do par ES-EN, foram extraídas por pivotagem (usando PT como língua pivô), e somente 37,5% foram extraídas pelas medidas de semelhança. A estes valores convém acrescentar o facto de a pivotagem e as medidas de semelhança utilizadas terem permitido uma extracção global com apenas 19% de falhas (81% de extracções correctas) para o par ES-EN [Tabela 22], em que as línguas são suficientemente distantes. Comparando com os pares PT-EN e PT-ES, para os quais não se extraiu traduções por pivotagem, e onde se obteve somente 64% e 50% de traduções válidas respectivamente, reconhece-se que o contributo da pivotagem para o enriquecimento do léxico é evidente, permitindo obter com maior fiabilidade, traduções de termos não cognatos e traduções de termos constituídos por múltiplas palavras.

Em resumo, este trabalho contribui com uma medida de semelhança nova (*Levenshtein*Dice*), que teve consequências positivas relativamente à precisão da extracção de traduções. A pivotagem, que não é propriamente uma novidade, acarretou extracções quer de traduções de palavras, quer de multi-palavras com um grau de precisão muito mais elevado, devido ao maior número de restrições introduzido, nomeadamente o posicionamento dos termos nos textos paralelos, e a existência de traduções confirmadas para o par PT-EN.
Ao estudar os três pares de línguas, PT, EN e ES, ficou claro que a extracção de traduções entre PT e ES é feita com muito menor precisão que a que se obtém dos pares EN-PT e EN-ES. Isto significa que, ao contrário do que é proposto por Mann e Yarowsky [Gideon, Mann et al., 2001] e por Shafer e Yarowsky [Shafer, Charles, 2002], há que ser extremamente cauteloso em extracções não supervisionadas. Esta conclusão mostra ainda que seria importante utilizar o inglês como língua pivô, para suportar melhor esta "descoberta" que é contudo natural. Entre línguas muito próximas é maior a possibilidade de confusão entre palavras individuais cognatas que não são tradução umas das outras.

Em trabalho futuro, uma vez que só se realizou a extracção de traduções por pivotagem para o par ES-EN, seria interessante efectuar também a extracção para os pares PT-EN e PT-ES, usando EN como língua pivô, e verificar se o respectivo léxico também usufrui de um incremento considerável no número de entradas válidas, à semelhança dos resultados obtidos para o par ES-EN.

O algoritmo criado pode ainda ser melhorado com a introdução de reconhecimento de padrões linguísticos entre segmentos de pares de línguas. Esta funcionalidade é conseguida com a utilização do extractor descrito em [Lopes, Gabriel e Aires, José, 2009], e que permitirá obter ainda mais traduções de multi-termos. Utilizando os pares de traduções extraídas, correctas e incorrectas, apesar de serem por enquanto em número bastante limitado, poder-se-á treinar uma *Support Vector Machine* para classificar novas traduções extraídas, apostando na sua capacidade para obter valores de precisão superiores a 95% quer na classificação das traduções correctas quer das incorrectas.[14]

Dado que os léxicos podem ser usados no refinamento de alinhamentos de textos, pode-se utilizar o alinhador descrito em [Gomes, Luís, 2009] para verificar qual a melhoria nos alinhamentos introduzida pelo enriquecimento do léxico. Esta experiência tem especial interesse, pois permitirá comparar directamente os resultados obtidos com o trabalho apresentado em [Bilbao, Darriba et al., 2005], onde apenas se recorre a possíveis cognatos extraídos, tendo em conta a medida de *Levenshtein Modificada* e a filtragem introduzida pelo processo de alinhamento.

[14] Kavitha Mahesh, na cadeira de MLKE, do3º Ciclo, realizou trabalho nesta direcção, ainda não publicado, e que apontava para valores de precisão e de *recall* daquela natureza, para o par EN-PT (comunicação pessoal)

5. Bibliografia

BILBAO, V. M. Darriba; LOPES, J. G. Pereira; ILDEFONSO, T. 2005. *Measuring the impact of cognates in parallel text alignment*. In: Carlos Bento, Amilcar Cardoso and Gael Dias (eds.) 2005: Portuguese Conference on Artificial Intelligence, Proceedings. Covilhã, December 2005. pp. 338-343. IEEE. ISBN 0-7803-9365-1.

BROWN, P. F.; LAY, J. C.; MERCER, R. L. 1991. *Aligning sentences in parallel corpora*. Proceedings of the 27 annual meeting of the ACL, Berkley, CA, pp 169-176.

CHURCH, K. W.; GALE, W. A. 1995. Poisson mixtures. *Natural Language Engineering*, 1(2). pp.163-190.

FRAKES, W. B.; BAEZA-YATES, R. (eds.). 1992. *Information Retrieval – Data Structures and Algorithms*. Englewood Cliffs, NJ: Prentice-Hall.

FUNG, Pascale. 1998. *A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-parallel Corpora*. In Proceedings of the Third Conference of the Association For Machine Translation in the Americas on Machine Translation and the information Soup (October 28 - 31, 1998). D. Farwell, L. Gerber, and E. H. Hovy, Eds. Lecture Notes In Computer Science, vol. 1529. Springer-Verlag, London, 1-17.

GALE, W. A; CHURCH, K. W. 1993. *A program for aligning sentences in bilingual corpora*. Computational linguistics, Vol. 19, 75-102.

GAMALLO, Pablo; CAMPOS, José Pichel. 2005. *An approach to acquire translations from non parallel texts*. In C. Bento, A. Cardoso and G. Dias (Editores). Progress in Artificial Intelligence. LNAI 3808 Springer Berlin.

GIDEON, S. Mann; YAROWSKY, David. 2001. *Multipath Translation Lexicon Induction via Bridge Languages*. Proceedings of the Second meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies (NAACL), pp.1-8.

GIGUET, Emmanuel; LUQUET, Pierre-Sylvain. 2006. *Multilingual Lexical Database Generation from parallel texts in 20 European languages with endogenous resources*. Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 271–278, Sydney, July 2006 Association for Computational Linguistics.

GOMES, Luís. 2009. *Parallel Texts Alignment*. Master Degree thesis, Universidade Nova de Lisboa, Lisboa.

HENDERSON, J.C. 2003. Word Alignment Baselines, HLT-NAACL 2003 Workshop: *Building and Using Parallel Texts Data Driven Machine Translation and Beyond*, pp. 27-30 Edmonton, May-June 2003.

HJELM, Hans. 2007. *Identifying Cross Language Term Equivalents Using Statistical Machine Translation and Distributional Association Measures*. Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007. Editors: Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit. University of Tartu, Tartu, 2007. ISBN 978-9985-4-0513-0 (online) ISBN 978-9985-4-0514-7 (CD-ROM) pp. 97-104.

HUTCHINS, W. John. 2006. *Machine translation, a concise history*. (http://www.hutchinsweb.me.uk/CUHK-2006.pdf)

JOACHIMS, Thorsten. 1998. *Text categorization with support vector machines: Learning with many relevant features*. In Proceedings of the European Conference on Machine Learning (ECML), 137-142, Springer.

KAY, Martin; RÖSCHEISEN, Martin. 1993. *Text-Translation Alignment*. In Computational Linguistics, volume 19, number 1, pp. 121–142.

LOPES, Gabriel; AIRES, José. 2009. *Phrase Translation Extraction from Aligned Parallel Corpora Using Suffix Arrays and Related Structures*, EPIA, LNAI 5816 Springer. p.587–597.

OAKES, Michael. 1998. *Statistic for Corpus Linguistics*. Edinburgh University Press, Edinburgh, Scotland, United Kingdom, 287 p.

OCH, F.; NEY, H. 2003. *A systematic comparison of various statistical alignment models*. Computational linguistics, Vol. 29(1), pp.19-51.

RIBEIRO, António. 2002. *Parallel Texts Alignment for Extraction of Translation Equivalents*. PhD thesis, Universidade Nova de Lisboa, Lisboa.

RIBEIRO, António; LOPES, Gabriel; MEXIA, João. 2000. *Linear Regression Based Alignment of Parallel Texts Using Homograph Words*. In Werner Horn (ed.) ECAI 2000: Proceedings of the 14th European Conference on Artificial Intelligence, Berlin, Germany, 2000 August 20-25. Vol. 54 Amsterdam, The Netherlands: IOS Press. pp. 446-450.

SALTON, Gerard; MCGILL, Michael. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, USA, p. 448.

SHAFER, Charles; YAROWSKY, David. 2002. *Inducting Translation Lexicons via Diverse Similarity Measures and Bridge Languages*. Proceedings of the 6th conference on Natural language learning (CoNLL)-Volume 20, ACL, pp. 1-7.

SILVA, J.F. da; DIAS, Gaël; GUILLORÉ, Sylvie; LOPES, J.G.P. 1999. *Using Local Maxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units*. In: P. Barahon, editor, Progress in Artificial Intelligence: 9th Portuguese Conference on AI, EPIA'93, Évora, Portugal, September 1999, Proceedings, pages 113-132, Springer-Verlag, Berlin, Germany, Lecture Notes in Artificial Intelligence, Vol. 1695. (http://www.sigmod.org/dblp/db/conf/epia/epia99.html#SilvaDGL99)

SMADJA, Frank; MCKEOWN, Kathleen; HATZIVASSILOGLOU, Vasileios. 1996. *Translation Collocations for Bilingual Lexicons: A Statistical Approach*. In Computational Linguistics, volume 22, number 1, pp. 1–38.

YAMAMOTO, M.; CHURCH, K.W. 2001. *Using Sufix Arrays to Compute Term Frequency and Document Frequency for All Substrings in a Corpus*, Computational Linguistics, 27(1), pages 1-30.