Vencendo a escassez de recursos computacionais. Carvalho: Tradutor Automático Estatístico Inglês-Galego a partir do corpus paralelo Europarl Inglês-Português.

Paulo Malvar Fernández
Area of Language Technology, **imaxin**|software
paulomal@gmail.com

José Ramom Pichel Campos
Area of Language Technology, **imaxin**|software
jrampichel@imaxin.com

Óscar Senra Gómez
Area of Language Technology, **imaxin**|software
oscar@imaxin.com

Pablo Gamallo Otero
Universidade de Santiago de Compostela
pablogam@usc.es

Alberto García
Igalia Free Software Company
agarcia@igalia.com

Resumo

À hora de desenvolver muitas ferramentas estatísticas de Processamento da Linguagem Natural torna-se essencial a utilização de grandes quantidades de dados. Para salvar a limitação da escassez de recursos computacionais para línguas minorizadas como o galego é necessário desenhar novas estratégias. No caso do galego, importantes romanistas têm teorizado que galego e português são variantes do português europeu. De um ponto de vista pragmático, esta hipótese poderia abrir uma nova linha de investigação para fornecer ao galego ricos recursos computacionais. Partindo do corpus paralelo inglês–português Europarl, **imaxin**|software compilou um corpus paralelo inglês–galego que utilizamos para criar um protótipo de tradutor automático estatístico inglês–galego, cuja performance é comparável a Google Translate. Mantemos que é possível implementar esta estratégia para desenvolver uma grande variedade de ferramentas computacionais para línguas, como o galego, intimamente relacionadas com línguas que já contam com um grande repertório de recursos computacionais.

1 Prefácio

Do ponto de vista da teoria lingüística sistémico-funcional hallidiana, as línguas funcionam, de acordo com Gee (1999, 1) "*tanto como uma ferramenta para a acção quanto como um andaime para as relações humanas dentro das culturas e grupos sociais e instituições*"[1]. Noutras palavras, a linguagem funciona como uma ferramenta não só para a comunicação mas para negociar as relações e as estruturas sociais da própria sociedade. É precisamente, mercê a esta dimensão social que a linguagem joga um papel simbólico crucial. Ao desenvolverem ferramentas computacionais para línguas concretas, os linguistas computacionais, sejam principalmente informáticos ou linguistas, são responsáveis para com as línguas com que trabalham. É possível que no caso de línguas prestigiadas esta responsabilidade não pareça óbvia. Nestes casos, as decisões a respeito de que fenómenos linguísticos se estudam e (mais importante do ponto de vista deste artigo) que ferramentas se desenvolvem; podem parecer triviais, pois semelham não implicar nenhum posicionamento ideológico. Porém, aqueles cientistas que decidiram trabalhar com e para línguas minorizadas, especialmente se são falantes dessas línguas, as suas decisões não são nunca inócuas.

É com esta responsabilidade como investigadores linguísticos e falantes que foi levado a cabo o projecto sobre o qual se debruça este artigo.

2 Introdução

Em 2008 e 2009, em **imaxin**|software levamos a cabo um projecto, subsidiado pola Dirección Xeral de I+D+i da Xunta de Galicia, chamado "RecursOpentrad: Recursos lingüístico-

[1] Tradução dos autores

computacionais para a tradución automática avanzada de código aberto para a integración europea da lingua galega". Dentro deste projecto, além de construirmos um sistema inglês–galego de Tradução Automática (TA) baseada em regras , pensamos que, dados os progressos[2] na actualidade atingidos no campo da Tradução Automática Estatística (TAE), era um excelente momento para dar mais um passo no desenvolvimento de ferramentas de Processamento da Linguagem Natural (PLN) para o galego.

Quando decidimos desenvolver um protótipo de um sistema de TAE inglês–galego, sabíamos que "*quanto maior [fosse] o corpus de treino disponível, mellor [seria] o desempenho [do] sistema de tradução*"[3] (Popović e Ney, 2006, 25) que poderíamos conseguir. Contudo, enquanto compilávamos os recursos necessários para o desenvolvimento de um protótipo para o citado par de línguas, chegamos à seguinte conclusão, absolutamente coincidente com uma das afirmações com que Popović e Ney (2006) começam a sua comunicação em Language Resources and Evaluation (LREC) em 2006:

"*Whereas the task of finding appropriate monolingual text for the language model is not considered as difficult, acquisition of a large high-quality parallel text for the desired domain and language pair requires a lot of time and effort, and for some languages is not even possible.*"(Popović e Ney, 2006, 25)

Convém termos em conta que não é impossível encontrar corpora paralelos inglês–galego na Internet.[4] De facto, o grupo de investigação de Xavier Gómez Guinovart na Faculdade de Tradução e Interpretação da Universidade de Vigo dispõe de uma colecção de corpora parlelos[5] dentro da qual o par inglês–galego está representado com um subcorpus de aproximadamente 9 milhões de palavras. Um corpus deste tamanho, porém, é a todos os efeitos insuficiente para o propósito de construír um sistema de TAE.

Chegados a este ponto, tornava-se, na nossa opinião, necessário tomar um rumo diferente para conseguirmos o nosso objectivo. É, neste sentido, conhecido na comunidade linguística que importantes romanistas, como por exemplo Coseriu (1987), Cunha e Cintra (2002) e Aracil (1985), têm teorizado que, de um ponto de vista linguístico, o galego deve ser considerado uma variante do português junto com o português europeu, brasileiro, africano e asiático. Isto é exactamente o que Coseriu (1987) e Rei (1991) apontam:

"*los romanistas e hispanistas están en general de acuerdo en que el gallego es una forma particular del conjunto dialectal gallego-portugués, en cuanto opuesto al conjunto dialectal español (no "castellano", sino: astur-leonés, castellano, en sus muchas formas, y navarro-aragonés) y al conjunto catalán (o catalán-valenciano)*" (Coseriu, 1987, 795)

"*Na actualidade, desde o punto de vista estrictamente lingüísitico, ás dúas marxes do Miño fálase o mesmo idioma, pois os dialectos miñotos e transmontanos son unha continuación dos falares galegos, cos que comparten trazos comúns que os diferencian dos do centro e sur de Portugal; pero no plano da lingua común, e desde unha perspectiva sociolingüísitca, hai no actual occidente peninsular dúas linguas modernas, con diferencias fonéticas, morfosintácticas e léxicas, que poden non impedi-la intercomprensión ó existir un bilingüismo inherente entre o galego e o portugués, semellante ó existente entre o catalán e o occitano, o danés e o noruegués, o eslovaco e o checo, o feroés e o islandés.*" (Rei, 1991, 17–18)

Deste modo, partindo da suposição de que galego e português são variantes linguísticas intimamente relacionadas e tentando aproveitar a posição privilegiada do português como língua computacionalmente desenvolvida –isto é, uma língua para a qual muitas ferramentas e recursos de PLN foram desenvolvidos–, em **imaxin**|software investigámos a possibilidade de utilizar corpora paralelos inglês–português de livre acesso para criar um corpus paralelo inglês–galego que utilizaríamos para desenvolver um protótipo de tradutor automático estatístico inglês–galego.

3 Compilação e processamento do corpus

3.1 O corpus de origem

Já que o nosso projecto estava claramente guiado pola filosofia do movimento do *Open Source*,

[2]Sirva de exemplo a grande popularidade da Tradução Automática Estatística (TAE) de alta qualidade atingida com a implementação feita por Google do seu sistema de TAE, Google Translate (disponível para consulta on-line em `http://translate.google.com/`).

[3]Tradução dos autores

[4]Graças à localização de projectos de ferramentas e sistemas operativos de código aberto levados a cabo pola comunidade galega de usuários de código aberto é possível compilar manualmente corpora paralelos inglês–galego do domínio da localização de software publicados baixo a General Public License (GPL). Contudo, estes corpora, ao serem traduzidos de maneira voluntária por grupos de pessoas não coordenados, não têm uniformidade e o seu tamanho resulta insuficiente para o propósito de crear um sistema de TAE.

[5]Esta colecção pode-se consultar em `http://sli.uvigo.es/CLUVI/`.

queríamos que tantos componentes do sistema como for possível fossem de código aberto, ou polo menos de livre acesso para uso não comercial.

Devido ao seu grande tamanho e liberal licença de *copyright*[6] escolhemos o corpus paralelo Europarl v3[7] inglês–português como corpus de origem do nosso projecto.

O corpus Europarl é um corpus paralelo extraído das Actas do Parlamento Europeu que inclui versões, desde 1996, do seu contido em onze línguas europeias: línguas romances (francês, italiano, espanhol e português), línguas xermánicas (inglês, neerlandês, alemão, danés e sueco), grego e finlandês.

Após um processo inicial de limpeza das etiquetas XML que marcam a estrutura discursiva das elocuções contidas no corpus, obtivemos um corpus paralelo inglês–português não-tokenizado que contém quase 65 milhões de palavras em total. Este corpus foi realinhado oração-a-oração[8] após o citado processo de limpeza empregando a ferramenta *sentence aligner*[9], incluída entre as ferramentas do Europarl v3.

3.2 Conversão de inglês–português a inglês–galego

A conversão do corpus paralelo de origem num corpus paralelo inglês–galego que desenhamos em **imaxin**|software é um processo semi-automatizado que envolveu o uso de duas peças de software principais: um sistema de tradução automática baseada em regras e um conversor ortográfico –isto é, um motor de transliteração.[10]

Deste modo, o fluxo de trabalho desenhado foi o seguinte:

- Tradução automática para galego do *lado* português do corpus paralelo de origem utilizando EixOpentrad.[11]
- Identificação dos erros de tradução devidos a erros de codificação de EixOpentrad. Quando em EixOpentrad existe uma regra de transferência ou uma entrada de dicionário mal formulada, o tradutor falha e marca a existência deste tipo de erros imprimindo os caracteres @ ou #, dependendo do tipo de erro, junto às palavras motivadoras dos erros.
- Revisão e correcção manual dos erros de tradução marcados com @ e #. As palavras marcadas com @ são palavras deficientemente codificadas no dicionário bilingue do tradutor. Os erros marcados com # corresponde-se, por sua vez, bem com erros de codificação nos dicionários monolingues, bem com erros de construção das regras de transferência do tradutor.
- Identificação das palavras desconhecidas, e portanto, não traduzidas por EixOpentrad. EixOpentrad marca as palavras não traduzidas com *, de modo que a sua identificação pode ser totalmente automatizada.
- Transliteração para galego das palavras desconhecidas, marcadas com *, utilizando um script de transliteração português–galego chamado port2gal.[12] As palavras que se transliteram no corpus são também armazenadas numa lista com a sua correspondente versão original não transliterada para a sua posterior revisão.
- Revisão e correcção manual dos erros de transliteração identificados na lista de palavras transliteradas obtidas no processamento anterior. Este processo de correcção, que não pode ser automatizado, é o passo que mais demora em se completar devido o tamanho limitado dos dicionários de EixOpentrad. É também, dada a sua extensão em número de palavras afectadas, um passo que convém realizar exaustivamente para assegurar a qualidade do corpus galego que se deseja obter.

[6] "The European Parliament web site states: "Except where otherwise indicated, reproduction is authorised, provided that the source is acknowledged."" (Koehn, 2005)(2)

[7] De livre acesso em `http://www.statmt.org/europarl/archives.html`

[8] Isto é, *sentence-to-sentece*.

[9] Esta ferramenta pode ser descarregada no site `http://www.statmt.org/europarl/v3/tools.tgz`

[10] Os conversores ortográficos utilizam-se normalmente para escrever o mesmo código de duas maneiras diferentes. Este tipo de conversores não fazem mais do que substituir padrões de sequeências de caracteres da língua de origem nos seus correspondentes padrões de sequeências de caracteres na língua de chegada. Esta estratégia não envolve informação morfológica, sintáctica nem semântica.

[11] EixOpenTrad é uma versão posterior de OpenTrad, uma plataforma de serviços de tradução automática (`http://www.opentrad.com`). EixOpenTrad é um protótipo de tradução automática galego-português e português–galego que contém 8.500 palavras em ambas as direcções. Este sistema está baseado no motor de tradução de Apertium espanhol–português, (Armentao-Oller et al., 2006).

[12] port2gal, que é um simples script de Perl, foi inicialmente desenvolvido por Alberto García (Igalia Free Software Company) e posteriormente melhorado por Pablo Gamallo (Departmento de Língua Espanhola da Universidade de Santiago de Compostela). Este script simplesmente converte a ortografia do portugês europeu para a ortografia actual do galego. port2gal está disponível baixo GPL em `http://gramatica.usc.es/~gamallo/port2gal.htm`.

Todo este processo conversão demorou três meses de trabalho de uma só pessoa a tempo completo (isto é, à volta de 3.600 horas) em total em se finalizar. Este é, sem dúvida, um período de tempo insignificante se comparado com o esforço em tempo e custos que suporia a compilação manual de um corpus inglês–galego deste tamanho.

3.3 O corpus final

Após finalizar o processo de conversão do corpus inglês–português obtivemos um corpus tokenizado inglês–galego composto de 34.715.016 tokens em inglês e 34.688.010 tokens em galego. Isto é, de aproximadamente 69 milhões de palavras, tamanho que é significativamente maior do que o tamanho do corpus citado na secção 2.

4 Tradução Automática Estatística

É comummente aceite por investigadores e profissionais da tradução que o principal desafio de todo o processo de tradução de uma língua para outra é basicamente encontrar um equilibrio entre a fidelidade com o significado expressado na língua de origem e a fluidez do texto equivalente na língua de chegada. De acordo com Jurafsky e Martin (2008, 875), "*Statistical MT is the name for a class of approaches that do just this by building probabilistic models of faithfulness and fluency and then combining these models to choose the most probable translation*". Assim, a melhor tradução $\hat{T}$ de uma frase de origem concreta S pode-se formalizar do seguinte modo:

$$\hat{T} = \text{argmax}_{\text{T}}\text{fidelidade}(T,S)\text{fluidez}(T) \quad (1)$$

Esta intuitiva definição informal da melhor tradução $\hat{T}$ pode ser matematicamente redefinida como a probabilidade condicional de uma possível tradução dada uma frase concreta da língua de origem:

$$\hat{T} = \text{argmax}_{\text{T}} P(T|S) \quad (2)$$

Utilizando a Regra de Bayes esta probabilidade condicional pode ser reescrita como:

$$\hat{T} = \text{argmax}_{\text{T}} \frac{P(S|T)P(T)}{P(S)} \quad (3)$$

Já que $P(S)$ não varia pois permanece constante para qualquer provável tradução T, $P(S)$ pode-se ignorar:

$$\hat{T} = \text{argmax}_{\text{T}} P(S|T)P(T) \quad (4)$$

Após a aplicação da Regra de Bayes podemos ver que, embora a nossa formalização intuitiva fazia a tradução T condicional na frase de origem S, a equação 4 fai a S de origem condicional na tradução T. Este modo inverso de formalizar problemas estatísticos, que é normal nos modelos conhecidos como *Noisy Channel*, tem a vantagem de que a a equação resultante pode ser perfeitamente paralelizada com a definição informal do problema de encontrar a melhor tradução $\hat{T}$:

$$P(S|T) = \text{fidelidade}(\text{T},\text{S}) \quad (5)$$

$$P(T) = \text{fluidez}(\text{T}) \quad (6)$$

4.1 Alinhamentos Palavra-a-Palavra

Nos anos 90 o grupo de investigação de IBM em Yorktown Heights (NY) começou a publicar algoritmos, Brown et al. (1990) and Brown et al. (1993), que, com relativo sucesso, utilizavam uma derivação bayesiana do modelo do *Noisy Channel* para construir tradutores automáticos estatísticos. A aproximação de IBM começava por estabelecer alinhamentos palavra-a-palavra entre frases alinhadas num corpus paralelo. Os alinhamentos palavra-a-palavra simplesmente formalizam a ideia de que existe um mapeamento explícito, embora não perfeito, entre as palavras das frases de origem e de chegada dos corpora paralelos. Seguindo a mesma aproximação do modelo do *Noisy Channel*, os algoritmos de alinhamento palavra-a-palavra modelam a probabilidade condicional de uma frase de origem S dada uma tradução T, alinhando palavra-a-palavra estas frases S e T:

$$P(S|T) = \sum_{A} P(S,A|T) \quad (7)$$

Noutras palavras, para um par concreto de frases alinhadas, S e T, a probabilidade condicional de S dada T encontra-se sumando todos os possíveis alinhamentos palavra-a-palavra A entre S e T.

Já que normalmente não há disponíveis corpora parlelos etiquetados à mão[13], é necessário utilizar um algoritmo para calcular as probabilidades de correspondências palavra-a-palavra utilizando a informação dada pola co-ocorrência de palavras num conjunto de frases paralelas. Para a realização desta tarefa normalmente utiliza-se o algoritmo conhecido como *Expectation Maximization* (EM).[14]

[13] De facto, seria muito caro em termos económicos e de recursos humanos etiquetar à mão as correspondências palavra-a-palavra em corpora paralelos do tamanho necessário para obter tradutores automátivos estatísticos de qualidade.

[14] Para uma explicação detalhada do funcionamento

4.2 TAE baseada em frases

Embora na TAE baseada em frases, em inglês *Phrase-based Statistical Machine Translation*, como qualquer outro sistema de TAE, a tradução se formalize com mesma equação 4 básica, os sistema de TAE baseada em frases são diferentes em termos daquilo que constitui a unidade de tradução básica. Assim, a principal intuição por trás deste tipo de TAE é que as palavras nem sempre são a melhor unidade de tradução pois a correspondência entre línguas normalmente não é $1:1$. Poder-se-ia argumentar que esta limitação foi superada polos sistemas de TAE baseada em palavras desde que o algoritmo de tradução de Brown et al. (1993) apresentasse um modelo de tradução conceitualmente preparado para tratar os alinhamentos $1:n$. Porém, os sistemas de TAE baseada em frases, dão mais um passo simplificando o problema ao converterem os alinhamentos de palavras em unidades de maior ordem, conhecidos como *frases*.[15] Assim, os sistemas de TAE baseada em frases, não realizam mapeamentos entre várias unidades, mas antes entre uma unidade e outra, embora de maior ordem que as palavras.

O modelo de TAE baseada em frases que se seguiu no desenvolvimento do nosso protótipo de TAE inglês–galego é o descrito em Koehn, Och e Marcu (2003).

5 Carvalho: sistema de TAE inglês–galego

Tal e como foi mencionado na secção 2, Carvalho é um protótipo de tradução automática estatística para o par de línguas inglês–galego. Carvalho foi treinado seguindo o paradigma da mencionada TAE beasada em frases. Para o seu treino três peças principais de software foram utilizadas:

- GIZA++[16]: GIZA++, originalmente desenvolvido durante o *John Hopkins University 1999 Summer Workshop*, é uma implementação de Och e Ney (2000) de todos os algoritmos de alinhamento palavra-a-palvra de IBM assim como do algoritmo HMM, acrónimo de Hidden Markov Models.[17]
- Moses[18]: Moses é a implementação de Koehn et al. (2007) da sua proposta de TAE baseada em frases feita em 2003, Koehn, Och e Marcu (2003). Moses utiliza os alinhamentos palavra-a-palavra aprendidos por GIZA++ para criar um modelo de tradução baseada em frases utilizado para determinar a melhor tradução $\hat{T}$ dada uma frase de origem S.
- SRILM[19]: SRILM, que pode ser utilizado livremente com fins não comerciais, é um modelizador de língua, isto é, uma ferramenta que aprende sequências de n-gramas, que servem para determinar a fluidez das traduções saintes de Moses e, deste modo, reordenar o ranking de traduções para finalmente determinar a tradução mais provável $\hat{T}$. SRILM foi treinado utilizando o texto completo do "lado" inglês ou galego, dependendo da direcção de traducção, do corpus de treino de GIZA++ e Moses.

5.1 Carvalho *vs.* Google Translate

Para exemplificar visualmente o sucesso que supôs a utilização do corpus paralelo inglês–galego obtido após o processamento descrito na secção 3.2 gostaríamos de mostrar dous exemplos de tradução; um realizado por Carvalho e outro por Google Translate[20], da seguinte frase, tirada da entrada da Wikipedia *Art*[21]:

Art is the process or product of deliberately arranging elements in a way that appeals to the senses or emotions. It encompasses a diverse range of human activities, creations, and modes of expression, including music, literature, film, sculpture, and paintings. The meaning of art is explored in a branch of philosophy known as aesthetics.

A tradução realizada por Carvalho é a seguinte:

Arte é o proceso ou produto de arranxar deliberadamente elementos dunha forma que apela á sentidos ou emocións. Engloba un diversificado abano de actividades humanas, creacións e modos de expresión, inclusive da música, da literatura, filmes, escultura e pinturas. O significado de arte é explotada en un ramo da filosofía coñecida como aesthetics.

À continuação mostra-se a tradução realizada por Google Translate a dia 2 de Março de 2010:

deste algoritmo ver Jurafsky e Martin (2008, 886–888).

[15] As *frases* na TAE baseada em frases não estão em absoluto linguisticamente motivadas, pois nada têm a ver com o conceito linguístico de frase derivado da teoria sintáctica de constituintes. Mesmo assim, empregaremos esta denominação pois é a mais estendida no campo da TAE.

[16] Disponível em `http://fjoch.com/GIZA++.html`.

[17] Para uma descrição detalhada do funcionamento dos algoritmos de IBM e HMM ver Och e Ney (2003).

[18] Disponível em `http://www.statmt.org/moses/`.

[19] Disponível em `http://www.speech.sri.com/projects/srilm/`

[20] O serviço de tradução de Google, Google Translate, incorporou em 2008 o galego entre o seu catálogo de línguas com ferramentas de PLN.

[21] `http://en.wikipedia.org/wiki/Art`.

	Inglês–Galego	Galego–inglês
Carvalho	0,1559	0,1895
GT	0,2559	0,3591

Tabela 1: Comparativa do *BLEU score* de Carvalho *vs.* Google Translate (GT).

A arte é o proceso ou produto de deliberadamente organizar elementos de un modo que pide aos sentidos ou emocións. Engloba unha variada gama de actividades humanas, creacións, e modos de expresión, incluíndo a música, literatura, cine, escultura e pintura. O significado da arte é explotado desde unha rama da filosofía coñecido como estética.

Embora resulte interessante poder comparar visualmente as traduções realizadas por estes dous sistemas de TAE, é obrigado empregar uma medida numérica objectiva para pôr em perspectiva a performance de ambos os sistemas. A medida escolhida foi *BLEU score*[22], (Papineni et al., 2001), calculada pola versão 11b do *National Institute of Standars and Technology* (NIST) dos Estados Unidos da América. A obtenção de um valor numérico de *BLEU score* realizou-se mediante a tradução de um pequeno corpus de refeência, *goldstandard*, de 11.500 palavras que compilámos em **imaxin**|software manualmente traduzindo uma colecção de 500 frases em inglês extraídas da versão online do jornal inglês *The Guardian*.

Em **imaxin**|software somos conscienteces de que as medidas obtidas têm as suas limitações. Por um lado, entre as críticas mais importantes vertidas a respeito de *BLEU score* está que esta medida em muito diversos contextos correlaciona-se deficientemente com as percepções humanas à hora de avaliar uma mesma tradução automática (ver Ananthakrishnan et al. (2007) ou Callison-Burch, Osborne e Koehn (2006)). Deste modo, se compararmos as traduções de exemplo de Carvalho e Google Translate, as diferenças entre dous sistemas parecem não ser tão dramáticas como sugerem os resultados obtidos mediante a tradução do nosso *goldstandard*, que apresentamos na tabela 5.1. Por outro lado, somos também conscientes de que, para a availiação de uma tradução automática, a utilização de uma só tradução de referência com *BLEU score* é insuficiente, já que se-lhe atribui a uma só tradução demasiado peso e valor, o qual não reflecte a realidade de que não existe uma *tradução perfeita* e que um mesmo texto de origem pode e deve ser traduzido de modos diferentes dependendo do contexto sociocultural, histórico, etc.

Tendo todas estas críticas em conta, a obtenção de uma medida numérica objectiva não deixa de ser útil quanto peça de informação de referência para comparar estes dous sistemas de TAE.

6 Conclusões

Neste artigo mostrou-se, por um lado, uma sólida estratégia de dramática redução do tempo de compilação de um corpus paralelo inglês–galego do tamanho necessário para o desenvolvimento de um protótipo de TAE para o citado par de línguas mediante o processo de conversão semi-automatizado descrito na secção 3.2. E demonstou-se, por outro lado, a alta qualidade dos resultados que podem ser obtidos seguindo esta estratégia (ver sec. 5.1). Estatégia que cremos foi também a seguida por Google na incorporação do galego no seu serviço Google Translate, tal e como sugere a seguinte tradução que em Abril de 2009 realizámos com este serviço durante os primeiros testes de avaliação de Carvalho:

A arte é o proceso ou produto de deliberadamente organizar elementos dun modo que apelido aos sentidos ou **emoções**. *Engloba un conxunto diversificado de actividades humanas,* **criações**, *e modos de expresión, incluíndo a música e a literatura. O significado da arte é explorador no ramo da filosofía coñecido como estética.*[23]

Em **imaxin**|software cremos firmemente que de não ser pola minimização do tempo de desenvolvimento e a alta qualidade dos resultados obtidos, Google muito provavelmente teria demorado muito mais tempo em incorporar o galego entre o leque de línguas das suas ferramentas de

[22] *BLEU score* é uma medida de avaliação de TA que mede a proximidade de uma tradução automática de uma tradução profissional humana, assumindo que quanto mais próxima esteja a tradução automática da tradução humana melhor é a primeira. Assim, o que *BLEU score*, a *grosso modo*, faz é contar o número de n-gramas da tradução automática que se sobrepõem aos da tradução humana, que se utiliza como tradução de referência. Na prática, *BLEU score* funciona combinando n-gramas sobrepostos ponderados de diferentes tamanhos –quatrogramas, trigramas, bigramas e unigramas. Além deste modelo de *backoff* de n-gramas sobrepostos, *BLEU score* também implementa um factor de penalização de brevidade que impede que as traduções sejam demasiado curtas com respeito à tradução humana de referência.

[23] Tal e como indica esta tradução, Google Translate foi muito provavelmente treinado utilizando corpora paralelos inglês–portugês parcialmente convertidos à ortografia galega. Contudo, à diferença da estrategia de **imaxin**|software, Google não parecia utilizar conversores ortográficos. Deste modo, as palavras portuguesas que não se encontravam nos seus dicionários permaneciam na sua ortografia original.

PLN.

É por tudo isto que podemos concluír com confiança que a estratégia de criar ferramentas de PLN para o galego partindo de recursos computacionais do portugês não é simplesmente justificável do poto de vista lingu ístico, mas absolutamente legítima.

Não é, do nosso ponto de vista, aventurado concluír que a utilização de recursos de uma língua intimamente relacionada, especialmente se esta é uma língua computacionalmente desenvolvida, é extremadamente útil para variedades linguísticas, como o galego, que carecem de ferramentas de PLN devido à sua posição de minorização.

Agradecimentos

A todos os investigadores/as que reconheceram que o galego tinha uma dimensão internacional e que tínhamos que nos aproveitar disso: Carvalho Calero, Manuel Rodrigues Lapa, Eugene Coseriu, etc.

Ao Parlamento Europeu por ter libertado as suas actas no domínio público.

À Dirección Xeral de I+D+i da Xunta de Galicia que financiou parte deste projecto RecursOpentrad.

Referências

Ananthakrishnan, R., P. Bhattacharya, M. Sasikumar, e R. M. Shah. 2007. Some issues in automatic evaluation of English–Hindi MT: More Blues for BLEU. Em *International Conference On Natural Language Processing (ICON)*.

Aracil, Ll. 1985. *Lingüística e sócio-lingüística galaico-portuguesa: reintegracionismo e conflito lingüístico na Galiza.* Asociaçom Socio-Pedagógica Galaico-Portuguesa.

Armentao-Oller, C., R. C. Carrasco, A. M. Corbí-Bellot, M. L. Forcada, M. Ginestí-Rosell, S. Ortiz-Rojas, J. A. Pérez-Ortiz, G. Ramírez-Sánchez, Felipe Sánchez-Martínez, e M. A. Scalco. 2006. Open-source Portuguese–Spanish machine translation. Em *Lecture Notes in Computer Science 3960 (Computational Processing of the Portuguese Language, Proceedings of the 7th International Workshop on Computational Processing of Written and Spoken Portuguese, PROPOR 2006)*, pp. 50–59. (c) Springer-Verlag.

Brown, P., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, e P. Roosin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85.

Brown, P., S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, e R. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Callison-Burch, C., M. Osborne, e P. Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. Em *Proceedings of the European Association for Computational Linguistics (EACL)*, pp. 249–256.

Coseriu, E. 1987. El gallego en la historia y en la actualidad. Em *Actas do II Congresso Internacional da Língua Galego-Portuguesa*, pp. 793–800.

Cunha, C. e L. Cintra. 2002. *Nova Gramática do Português Comtemporâneo.* Edições João Sá da Costa.

Gee, J. P. 1999. *An Introduction to Discourse Analysis: Theory and Method.* Routledge.

Jurafsky, D. e J. H. Martin, 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, chapter 25, pp. 859–908. Pearson, 2 edition.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. Em *MT Summit 2005.*

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, M. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, e E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. Em *Annual Meeting of the Association for Computational Linguistics (ACL).*

Koehn, P., F. J. Och, e D. Marcu. 2003. Statistical phrase-based translation. Em *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pp. 127–133.

Och, F. J. e H. Ney. 2000. Improved statistical alignment models. Em *Proceedings of 38th Annual meeting of the ACL*, pp. 400–447.

Och, F. J. e H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Papineni, K. A., T. Roukos, T. Ward, e W. J. Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. Relatório técnico, IBM Research Division, Thomas J. Watson Research Center.

Popović, M. e H. Ney. 2006. Statistical machine translation with a small amount of bilingual training data. Em *Language Resources and Evaluation (LREC) 5th SALTMIL Workshop on Minority Languages: "Strategies for developing machine translation for minority language"*, pp. 25–29.

Rei, F. Fernández. 1991. *Dialectoloxía da lingua galega*. Edicións Xerais de Galicia.