

O Problema da Ambigüidade Lexical de Sentido na Comunicação Multilíngüe

Lucia Specia, Maria das Graças Volpe Nunes

Instituto de Ciências Matemáticas e de Computação – USP
Av. do Trabalhador São-Carlense, 400, Caixa Postal 668 – 13.560-970 – São Carlos – SP
{lspecia,gracan}@icmc.usp.br

Abstract. *In this paper it is presented a discussion about the problem of word sense ambiguity in computational systems aiming at multilingual communication, specially at machine translation from English into Brazilian Portuguese. In order to do this, examples of sentences showing the implications of this linguistic phenomenon and the way it is addressed by different machine translators are analyzed. The goal is to demonstrate the importance of such module in an English-Portuguese multilingual communication system and then justify the purpose of a word sense disambiguation model between these two languages.*

Resumo. *Neste artigo é apresentada uma discussão sobre o problema da ambigüidade lexical de sentido em sistemas computacionais voltados para a comunicação multilíngüe, em especial, para a tradução automática do inglês para o português do Brasil. Para tanto, são analisados exemplos de sentenças que apontam as implicações desse fenômeno lingüístico e o tratamento a ele dispensado por diferentes tradutores automáticos. Pretende-se, com isso, demonstrar a importância de um módulo dessa natureza em um sistema de comunicação multilíngüe inglês-português, para então justificar a proposta de um modelo de desambiguação lexical de sentido entre essas duas línguas.*

1. Introdução

A comunicação multilíngüe tem se tornado uma tarefa cada vez mais imperativa no cenário atual de grande disseminação de informações em diversas línguas, especialmente por meio da Internet. Nesse contexto, são de grande relevância os sistemas computacionais que auxiliam tal comunicação, automatizando-a, agregando a ela velocidade e praticidade. Dentre esses sistemas estão os de tradução automática, de recuperação de informações multilíngües, de categorização de textos multilíngües, etc.

Este trabalho está focalizado na Tradução Automática (TA), uma vez que essa aplicação representa uma etapa fundamental para a concretização de outras aplicações de comunicação multilíngüe, as quais utilizam, em algum momento, um módulo de tradução. Por exemplo, um sistema de recuperação de informações multilíngües normalmente emprega um módulo dessa natureza para identificar a tradução correta de um dado termo de busca nas diversas línguas envolvidas nessa busca e, com base nessa tradução, recuperar os documentos, em todas as línguas, que estão relacionados ao termo.

Muito embora seja uma das áreas mais antigas do Processamento de Línguas Naturais (PLN), a TA ainda apresenta muitos problemas. A maior parte desses problemas

está relacionada à ambigüidade entre as línguas. A ambigüidade interlingual ocorre também na tradução humana, mas é um problema especialmente grave na TA, uma vez que não se pode contar, nesse caso, com o conhecimento e a experiência da língua e do mundo do interpretador humano. A ambigüidade na tradução pode ser verificada nos diversos níveis de interpretação das línguas, incluindo o lexical, sintático, semântico, contextual e pragmático.

Este trabalho concentra-se na ambigüidade no nível lexical, que ocorre quando da multiplicidade de opções, durante a seleção de uma palavra, na língua-alvo (LA), para traduzir uma palavra da língua-fonte (LF). Ambigüidades dessa natureza caracterizam sempre uma escolha imprescindível e cujos efeitos podem ser extremamente prejudiciais à tradução, uma vez que diferentes opções podem dar origem a proposições semanticamente muito distintas. Esse problema se mostra ainda mais grave e de solução mais complexa quando são identificadas apenas variações de significado (de sentido) nas opções de tradução na LA, ou seja, quando essas opções são todas da mesma categoria gramatical. Essa variação do problema é chamada de ambigüidade lexical de sentido (ALS) e representa o foco deste trabalho, em oposição à ambigüidade categorial. A área que se ocupa do seu tratamento é denominada Desambiguação Lexical de Sentido (DLS), do inglês *Word Sense Disambiguation*.

A ALS é causada, fundamentalmente, pela existência de algumas relações semânticas interlexicais entre as línguas, principalmente a homonímia e a polissemia. Em uma das possíveis distinções entre essas duas relações, considera-se que, na **polissemia**, para uma determinada palavra da LF, existem duas ou mais palavras correspondentes na LA, com diferentes significados relacionados entre si. Por exemplo, à palavra do inglês *know* podem corresponder pelo menos duas palavras relacionadas no português, “saber” e “conhecer”. Já na **homonímia**, para uma dada palavra da LF correspondem duas ou mais palavras da LA, com diferentes significados, mas não relacionados entre si. Por exemplo, a palavra do inglês *light* pode ser traduzida como “leve” ou “luz”, entre outras opções. Neste trabalho, tal diferenciação não é relevante, pois pretende-se analisar ambos os fenômenos indistintamente.

Para ilustrar como a ALS é comum na tradução, considere a sentença abaixo e a quantidade de possíveis traduções (para o português) de cada uma das suas palavras de conteúdo, indicada entre parênteses, com base apenas nas traduções denotativas encontradas no dicionário eletrônico DTS DIC Prático Michaelis® 5.1. São consideradas, aqui, somente as traduções já na categoria gramatical adequada, ou seja, não há ambigüidade categorial.

“I expect (7) some (3) take (110) the veil (8) simply to hide (5) a flat (24) chest (11)”.

A média de possíveis traduções para as palavras analisadas dessa sentença é 24. Se for considerado, ainda, que o significado de cada palavra pode depender do significado das demais palavras na sentença, o número de combinações possíveis é de aproximadamente 24⁷.

Para realizar a desambiguação de maneira automatizada, os sistemas de TA devem incorporar um módulo de DLS ao processo de tradução. Várias abordagens têm sido propostas para a criação de módulos de DLS. Contudo, essas abordagens, na sua maioria, não são empregadas na TA, mas sim em aplicações monolíngües, as quais apresentam características bastante distintas das multilíngües no que se refere à manipulação da ambigüidade. Em se tratando de ambientes multilíngües envolvendo o português, em especial, não se tem conhecimento de módulos de DLS desenvolvidos e efetivamente empregados.

O objetivo deste trabalho é mostrar que a ALS é um problema grave para a

comunicação multilingüe e como a falta de mecanismos de DLS afeta negativamente essa comunicação, considerando a TA inglês-português como cenário. Para tanto, são apresentados três estudos realizados com base em diversos tradutores inglês-português e sentenças de diferentes gêneros e domínios. Atenção especial é dispensada à investigação do problema da ALS na tradução de verbos, entretanto, também são apresentados resultados da análise do problema em outras classes de palavras. Este trabalho também realiza uma investigação sobre as pesquisas na área de DLS para a comunicação multilingüe, incluindo a TA, visando mostrar que não há propostas efetivas, ainda que teóricas, envolvendo a língua portuguesa.

A partir da análise do problema da ALS nos diversos sistemas de TA e da constatação de que não há abordagens para a DLS multilingüe que considerem a língua portuguesa, pretende-se, posteriormente, justificar e embasar a proposta um modelo de DLS a ser empregado em sistemas de TA do inglês para o português do Brasil.

Para mostrar as implicações do problema da ALS na tradução e o comportamento dos sistemas de TA diante desse fenômeno, na Seção 2 são apresentados os três estudos sobre a ocorrência desse problema nos diferentes sistemas. As abordagens existentes para o problema da ALS voltadas para a comunicação multilingüe são ilustradas na Seção 3. Algumas considerações e possíveis direcionamentos desse trabalho são discutidos na Seção 4.

2. O problema da ALS na TA

Os problemas causados pela ALS na tradução envolvendo o português do Brasil foram recentemente analisados em três estudos experimentais. O primeiro estudo, com base no qual este trabalho está especialmente fundamentado, consistiu da realização de um experimento com o *cópus BNC (British National Corpus)* (Burnard 2000) com o objetivo de investigar as conseqüências da ALS em traduções automáticas de textos reais, a fim de delimitar a proposta de um modelo de DLS aos casos mais problemáticos de ambigüidade¹. Esta atividade foi desempenhada com base em três sistemas de TA inglês-português comumente utilizados, a saber, Systran, FreeTranslation e Globalink Power Translator Pro. Foram considerados para análise somente os verbos das sentenças, inicialmente, o subconjunto dos 15 verbos mais freqüentes do BNC². Para a análise, 531 sentenças do BNC contendo esses 15 verbos foram aleatoriamente selecionadas e submetidas aos tradutores. As traduções foram, então, manualmente analisadas para verificar a ocorrência da ALS, seus efeitos na tradução das sentenças e o comportamento dos sistemas diante desse fenômeno.

Nesse estudo foram definidos critérios específicos para identificação de um subconjunto de verbos mais problemáticos com relação à ocorrência de ALS e à ineficiência no tratamento dispensado a ela pelos sistemas de TA. Com base nesses critérios foram selecionados sete verbos: *to go*, *to get*, *to make*, *to take*, *to come*, *to look* e *to give*. Alguns exemplos de casos de ALS encontrados no uso desses verbos e não manipulados adequadamente pelos tradutores avaliados são ilustrados na Tabela 1.

A partir desse estudo, pôde-se perceber que a porcentagem de sentenças nas quais

¹ Esse experimento é apresentado com detalhes em Specia and Nunes (2004).

² Essa categoria gramatical foi escolhida porque os verbos são altamente ambíguos e porque da sua desambiguação pode depender a desambiguação de outras palavras da sentença, principalmente dos seus argumentos. Posteriormente, pretende-se estender esse trabalho a outras categorias lexicais.

ocorre ambigüidade nos sete verbos selecionados é bastante grande. De acordo com os critérios definidos, foram consideradas sentenças problemáticas somente aquelas cuja acepção do verbo em foco não era corretamente identificada por pelo menos dois sistemas. Com base nesse critério, das 238 sentenças com os sete verbos, 149 foram consideradas problemáticas (62,6% do total). Se fossem consideradas problemáticas as sentenças nas quais a acepção correta do verbo não havia sido identificada por pelo menos um sistema, esse número aumentaria para 177 sentenças (74,4% do total). Esse número alto mostra que os sistemas estudados não dispõem de mecanismos de DLS. Normalmente, eles escolhem uma das possíveis acepções de um verbo, provavelmente a mais comum, e essa acepção é utilizada na tradução da maioria das suas ocorrências, excetuando-se alguns casos do uso do verbo em *phrasal verbs* ou em expressões comuns. O tratamento dispensado a *phrasal verbs* é também bastante simplificado: muitas vezes, um verbo seguido de uma preposição (dois elementos que poderiam compor um *phrasal verb*) é diretamente traduzido como o *phrasal verb* correspondente, mesmo que não seja usado com tal função na sentença, como ocorre com a última sentença da Tabela 1.

Tabela 1. Exemplos de sentenças do BNC com verbos problemáticos

Sentença	Acepção correta	TA		
		Systran	Free-Translation	Power Translator
The war may well just go on and on.	continuar	ir	vai	ir
Stand in a French village when the Tour de France goes by and you are participating in an event which is unambiguously French.	passa (passar)	vai	vai	passa
It's best to be alone when the noises get this loud.	ficam (ficar)	recebem	começam	adquirem
A lot of international help will be needed to get things moving.	fazer	receber	começar	adquirir
They take more foreign holidays.	têm (ter)	tomam	fazem exame	levam
" Take that money out of your mouth!" said her mother.	tire (tirar)	toma ... fora	faça exame ... fora	objeto pegado ... fora
Now eat your supper, both o' ye, afore it takes cold.	fique (ficar)	toma	faz exame	leva
"This city has suddenly come alive," said her husband, an off-duty border guard.	renasceu (renascer)	veio vivo	vivo ... vindo	veio viva
"Yes, I'm coming , but I've one or two things to attend to first," she explained.	indo (ir)	venho	vindo	vindo
Mr Gonzalez has also come in for criticism from within his own party.	recebeu (receber)	entrou	entrou	entrou

O segundo estudo (Fossey et al. 2004) foi desenvolvido como parte da avaliação do sistema de TA inglês-português EPT-Web³, ainda em construção. Nele, foi analisada, entre outros problemas, a ocorrência da ambigüidade lexical na TA inglês-português. Para tanto, foi considerado um corpus de textos do jornal *New York Times* (NYT) *on-line* e quatro ferramentas de tradução disponíveis na web: Linguatrec E-Translation Server, Intertran, Systran e FreeTranslation. Esse estudo considerou indistintamente ambos os tipos de ambigüidade lexical, isto é, categorial e de sentido (homonímia e polissemia). Foram avaliadas

³ <http://www.nilc.icmc.usp.br/nilc/projects/ept-web.htm>

as traduções de todas as palavras de conteúdo de 515 sentenças, nos quatro sistemas. Uma sentença foi considerada problemática em um sistema se apresentasse pelo menos uma palavra ambígua inadequadamente traduzida por esse sistema. Os números e percentuais de sentenças problemáticas, em cada tradutor, são apresentados na Tabela 2.

Tabela 2. Sentenças do NYT com ambigüidade lexical

Sistema	Nº de sentenças cuja acepção não foi corretamente identificada	% de sentenças cuja acepção não foi corretamente identificada
E-Translation	279	54,1
Intertran	361	70,1
Systran	272	52,8
FreeTranslation	271	52,6

No estudo também foram apresentados os percentuais de palavras ambíguas cuja acepção não foi corretamente identificada pelos tradutores, agrupadas de acordo com a sua categoria gramatical, com relação ao total de palavras ambíguas. Nos quatro sistemas avaliados, a maioria das palavras ambíguas se distribuía entre substantivos e verbos, conforme ilustrado na Tabela 3. Exemplos de sentenças com problemas de tradução causados pela ALS apenas dos verbos em alguns sistemas são ilustrados na Tabela 4.

Tabela 3. Verbos e substantivos ambíguos do NYT

Sistema	% de substantivos cuja acepção não foi corretamente identificada	% de verbos cuja acepção não foi corretamente identificada
E-Translation	36,7	29,8
Intertran	38,7	32,3
Systran	39,6	24,1
FreeTranslation	40,0	31,4

Tabela 4. Exemplos de sentenças do NYT com ALS nos verbos

Sentença	Acepção correta	TA	Sistema de TA
With an Organic Sensor, a Food Wrapper Sniffs Out Trouble.	descobre (descobrir)	funga fora	E-Translation
Bush Sending Powell to Middle East.	enviando (enviar)	emite	Systran
Click Here to Receive 50% Off Home Delivery of The New York Times Newspaper.	clique (clicar)	estale	Systran
Check them out , or post any wine-related topics.	dê baixa (dar baixa)	verifique-os para fora	FreeTranslation

Pela Tabela 2, pode-se observar que os quatro sistemas apresentaram um porcentual maior que 50% de sentenças com problemas específicos de ambigüidade lexical, em uma ou mais palavras. Em uma análise realizada, ainda no estudo citado, sobre a gramaticalidade das sentenças, foi verificado que a maior parte de sentenças problemáticas correspondem a sentenças agramaticais ou gramaticais com tradução incorreta. Com isso, em alguns casos, ainda que os problemas de ambigüidade lexical fossem resolvidos, as sentenças permaneceriam incorretas, já que eram agramaticais, mas, em grande parte dos casos, uma vez resolvido o problema da ambigüidade, as sentenças poderiam, em sua maioria, tornar-se

semanticamente corretas. Assim, segundo os autores, fica evidente que a ambigüidade lexical compromete profundamente a qualidade das traduções produzidas automaticamente e que a solução das questões envolvendo esse problema se mostra um dos caminhos necessários para a obtenção de resultados mais satisfatórios nas produções das ferramentas de TA.

O terceiro estudo foi realizado por Oliveira et al. (2000). Os autores analisaram, comparativamente, vários sistemas de TA entre inglês e português, comerciais ou disponíveis na web, avaliando ambas as direções da tradução. Para testar o desempenho dos sistemas na direção inglês-português, 10 passagens de textos do jornal *New York Times* (com uma ou mais sentenças, totalizando 530 itens lexicais) foram submetidos a cinco sistemas: Globalink Power Translator Pro, Alta Vista, Intertran, Tradunet e Linguatex E-Translation Server. As traduções foram analisadas para identificar problemas em três níveis de interpretação: lexical, sintático e semântico-pragmático.

No nível lexical, o desempenho dos sistemas foi testado em quatro situações: dicionarização, ambigüidade, conotação e expressões idiomáticas. No caso das ambigüidades lexicais, foram consideradas, indistintamente, as ambigüidade categorial e de sentido (polissemia ou homonímia). Alguns exemplos de problemas causados pela ALS na tradução, selecionados entre os relatados, são ilustrados na Tabela 5.

Tabela 5. Exemplos de sentenças com ALS (Oliveira et al. 2000)

Sentença	Acepção correta	TA	Sistema de TA
(...) Hungary has ceded more sovereignty than many other nations – including the United States – would ever consider (...)	sempre	jamais	Translator Pro
To paraphrase a celebrated epitaph, prosperity left scarcely any of our industries untouched, and touched nothing it did not enrich.	mal	quase nenhuma	E-Translation

Segundo os autores, a presença da ambigüidade lexical na TA entre o inglês e o português é bastante freqüente, justificando a necessidade de estratégias de desambiguação nas ferramentas de tradução. Eles afirmam que a qualidade das escolhas lexicais afeta o processo de tradução em vários graus, principalmente se a escolha incorreta ocorrer em itens lexicais em posições de núcleo, como verbos em um predicado verbal ou substantivos em um sujeito. Nesses casos, a ambigüidade lexical pode prejudicar a coerência local e global da sentença, freqüentemente tornando-a incompreensível.

Nesse estudo, os autores também verificaram que as ferramentas de tradução não empregam mecanismos para procurar resolver o problema da ambigüidade lexical. Em vez disso, apostam em decisões baseadas em critérios muito simples, como a freqüência da ocorrência de cada acepção em traduções reais. A maioria dos erros encontrados, segundo os autores, diz respeito a expressões com grupos de palavras que podem assumir significados diferentes da composição do significado que elas possuem individualmente, como ocorre, por exemplo, em *phrasal verbs*. A conclusão geral dos autores é que a qualidade das traduções poderia ser consideravelmente aprimorada se fosse assumida uma perspectiva diferente com relação às idiossincrasias de cada língua, ou seja, se fossem empregados esforços de caráter mais efetivo para o tratamento dessas idiossincrasias.

De modo geral, apesar de terem objetivos distintos, os três estudos citados

apresentam resultados que corroboram a hipótese de que a ambigüidade lexical influencia negativamente nos resultados da comunicação multilingüe, em especial, na TA inglês-português, e que mostram que esse problema não recebe, ainda, tratamento adequado nas ferramentas disponíveis. Com isso, comprovam a necessidade de mecanismos de DLS para essa comunicação.

3. Abordagens para a DLS na comunicação multilingüe

Várias abordagens de DLS têm sido propostas para diversas aplicações, principalmente para aquelas monolingües. Essas abordagens podem seguir diferentes métodos de PLN: **métodos lingüísticos**, baseados em conhecimento lingüístico e/ou extralingüístico explicitamente especificado, manualmente ou semi-automaticamente, por meio de recursos como dicionários eletrônicos; **métodos empíricos**, baseados em *cópus* de exemplos e em algoritmos de aprendizado de máquina para adquirir conhecimento automaticamente a partir dos exemplos; ou **métodos híbridos**, que combinam características dos métodos lingüísticos e empíricos.

Considerando-se a aplicação específica da DLS em tarefas multilingües, são poucos os trabalhos desenvolvidos de que se tem conhecimento. No caso de abordagens que seguem métodos lingüísticos, pode-se citar os trabalhos de Egedi et al. (1994), Dorr and Katsova (1998), Pedersen (1997) e Montoyo et al. (2002). Egedi et al. (1994) apresentam um sistema de TA do coreano para o inglês que possui um módulo de DLS para tratar da polissemia de alguns verbos, com base na unificação de restrições de seleção semânticas definidas na estrutura argumental desses verbos com os traços semânticos definidos para os substantivos que podem ser utilizados como seus argumentos. As restrições de seleção e traços semânticos são especificados na LA. Assim, a desambiguação de um verbo depende da tradução correta dos seus argumentos. Certamente, tal abordagem apresentará problemas se os argumentos do verbo também forem ambíguos.

Dorr and Katsova (1998) definem um mecanismo de seleção lexical para verbos e substantivos deverbais que se baseia na estrutura argumental desses elementos, representada por meio de Estruturas Conceituais Lexicais, e nos sentidos da WordNet. A hipótese é de que a tradução de um elemento da LF pode ser desambiguada se forem escolhidos, na LA, elementos que apresentem a mesma LCS e que estejam no mesmo *synset* da WordNet, ou seja, que sejam sinônimos do elemento na LF. Em experimentos considerando a desambiguação do inglês para o espanhol, as autoras obtiveram resultados promissores, constatando que os elementos são facilmente desambiguados, pois são raros os elementos com a mesma LCS que são sinônimos. Contudo, o mecanismo proposto exige uma base de dados com todos os itens lexicais representados por estruturas LCS e previamente mapeados (manualmente) em um *synset* da WordNet.

Montoyo et al. (2002) discutem a necessidade de um módulo de DLS em aplicações multilingües voltadas para recuperação de informações, e apresentam uma interface para a desambiguação de substantivos e verbos que poderia ser acoplada a esses sistemas. Nessa interface, consideram o espanhol e o inglês como LF, e a taxonomia da EuroWordNet para realizar o mapeamento entre as palavras dessas duas línguas e também o mapeamento para o catalão e o basco. A desambiguação realizada consiste, basicamente, em identificar qual é o código da EuroWordNet correspondente à palavra a ser desambiguada, ainda na LF e, em seguida, encontrar a palavra na LA com o mesmo código da EuroWordNet. Assim, embora seja voltada para aplicações multilingües, a desambiguação é feita de maneira monolingüe.

Essa abordagem só se mostra viável para línguas previstas no projeto EuroWordNet, para as quais já existem códigos correspondentes aos itens lexicais.

Pedersen (1997) descreve um mecanismo para a desambiguação de um subconjunto de verbos de movimento polissêmicos na TA do dinamarquês para o inglês. A autora considera apenas a polissemia sistemática desse subconjunto. Para tanto, utiliza a abordagem de esquemas para descrever os verbos a partir de uma grande quantidade de informações lingüísticas da LF, em diversos níveis, para auxiliar na desambiguação. Apesar da aplicação voltada para a TA, o foco da autora é na especificação desses esquemas com informações suficientes para permitir capturar os padrões sistemáticos entre os diferentes sentidos de um verbo, de modo a evitar descrições ambíguas. Assim, a desambiguação ocorre, em grande parte, na LF. Além disso, a estrutura definida para os esquemas é específica para verbos de movimento do dinamarquês, o que dificulta extensões dessa abordagem.

No caso de abordagens que seguem métodos empíricos, podem ser considerados os trabalhos de Brown et al. (1991) e Lee (2002). Brown et al. (1991) descreve uma abordagem estatística para a seleção lexical na TA entre o francês e o inglês. Essa abordagem é bastante simples, pois deriva do teorema de Bayes, que se baseia principalmente na frequência de cada possível tradução em um corpus. Além disso, ela considera uma desambiguação binária, ou seja, a escolha entre apenas dois possíveis sentidos de uma palavra ambígua. O módulo de desambiguação desenvolvido foi avaliado em um sistema de TA, também estatístico, considerando as 500 palavras mais comuns do inglês e as 200 mais comuns do francês. Com esse módulo, a taxa de erro nas traduções resultantes do sistema, segundo os autores, diminuiu 13%. Contudo, é preciso levar em conta as características limitadas do contexto de desenvolvimento desse modelo.

Lee (2002) apresenta uma abordagem baseada em corpus para a seleção lexical na TA inglês-coreano. Essa abordagem trata a seleção lexical como um problema de classificação e emprega um algoritmo para a escolha pela classe (tradução) mais adequada. Tal algoritmo também é estatístico, mas contempla outras características, além da frequência das traduções. O autor tem por objetivo obter um modelo portátil, independente de língua. Para tanto, utiliza como características para a classificação somente as outras palavras da sentença em que a palavra ambígua ocorre, agrupadas de duas a duas. Com isso, a precisão da classificação gerada, quando o modelo é avaliado com novos casos, é pouco melhor que a da *baseline* considerada (neste caso, a tradução mais frequente). O autor não cita a abrangência do modelo, mas como esse modelo se baseia nas palavras da sentença, tal abrangência provavelmente é bastante limitada.

Considerando-se as abordagens de DLS para a TA que seguem métodos híbridos, pode-se citar o trabalho de Zinovjeva (2000). Esse trabalho tem por objetivo aprender automaticamente regras de transformação para traduzir corretamente palavras ambíguas do inglês para o sueco. Para tanto, utiliza conhecimento pré-codificado em recursos lingüísticos como dicionários, por meio de procedimentos, tais como um etiquetador morfossintático, que atuam como filtros, eliminando alguns dos possíveis sentidos de cada palavra ambígua. O resultado é um conjunto reduzido de sentidos para cada palavra ambígua, em um determinado contexto. Para a geração do modelo de DLS, são fornecidos exemplos de desambiguação a um algoritmo de aprendizado de máquina, baseado no método de aprendizado por transformações. Esses exemplos são formados por sentenças com palavras ambíguas, seu contexto e as suas correspondentes traduções, identificadas manualmente.

Gerado o modelo, para desambiguar um novo caso, os resultados dos procedimentos preliminares são fornecidos ao modelo, que escolhe pela tradução mais adequada. Apesar de ser considerada uma abordagem híbrida, a utilização do conhecimento pré-codificado e a aquisição de novos conhecimentos (ou seja, das regras de transformação) ocorrem em etapas isoladas no processo de DLS, assim, pouco se aproveita de todo o potencial das metodologias híbridas.

Os trabalhos citados, apesar de não representarem um lista exaustiva de todos os encontrados, exemplificam a maioria deles, mantendo a proporção da distribuição entre os tipos de abordagens. Como se pode perceber, a maioria desses trabalhos é baseada em métodos lingüísticos. Apesar de poderem apresentar resultados bastante precisos, as dificuldades para a criação das fontes de conhecimento acabam restringindo muito a abrangência desses trabalhos. Abordagens baseadas em métodos empíricos permitem modelos mais abrangentes, mas são ainda pouco pesquisadas no contexto da TA, em função da dificuldade na criação de corpus de exemplos representativos e consistentes. Abordagens seguindo métodos híbridos, por sua vez, permitem combinar as vantagens de ambos os métodos, mas são raros os trabalhos desenvolvidos sob esse método.

Pode-se observar, também, que nenhum dos trabalhos citados inclui desambiguações para o português. No único trabalho envolvendo o português encontrado, Leffa (1998) cita a importância do uso do contexto local da palavra ambígua, isto é, das palavras vizinhas a ela na sentença, para a desambiguação na TA. Em um experimento para desambiguar 20 palavras ambíguas do inglês para o português, o autor relata um desempenho de 94%. No entanto, como esse modelo se baseia nas palavras da sentença, sua abrangência deve ser bastante limitada. Além disso, o trabalho foi interrompido.

4. Considerações finais

Neste trabalho foram relatados os resultados de alguns estudos experimentais com diversos sistemas de TA inglês-português, focalizando o problema da ALS. Por meio desses estudos foi possível constatar que a ALS é um problema bastante comum, proeminente e prejudicial para a tradução. Além disso, foi possível verificar que todos os sistemas testados, de uso expressivo atualmente, não oferecem tratamento adequado para esse problema. Em se tratando da TA para o português, particularmente, de fato, não se tem conhecimento de sistemas que empreguem mecanismos de DLS. Pode-se concluir, com isso, que a falta desses mecanismos é certamente um dos principais motivos para os resultados bastante insatisfatórios dos sistemas existentes.

Considerando-se as pesquisas teóricas, foi possível verificar que há poucos trabalhos especificamente voltados para a TA e que as metodologias normalmente exploradas não permitem criar abordagens com resultados precisos e, ao mesmo tempo, abrangentes. Além disso, não se tem conhecimento de trabalhos significativos envolvendo o português.

A partir dos resultados dessa análise sobre o problema da ALS na TA inglês-português e da investigação das abordagens existentes para esse problema (para outras aplicações ou envolvendo outras línguas), pretende-se desenvolver um modelo computacional de DLS voltado especificamente para a TA do inglês para o português do Brasil. Esse modelo será construído seguindo um método realmente híbrido de PLN, ou seja, será baseado em conhecimento lingüístico e em corpus, com a utilização de conhecimento substancial durante o processo de aprendizado automático. Essa configuração permitirá

que o modelo seja abrangente, aplicável a sistemas de TA em larga-escala, auxiliando no processo de escolha lexical de uma grande quantidade de palavras, e que apresente resultados potencialmente melhores que os dos trabalhos já existentes para outras línguas. Adicionalmente, esse modelo deverá ser o mais independente possível, de maneira que possa ser acoplado a diferentes sistemas de TA.

Desenvolvida uma abordagem eficiente para o problema da ALS, esta poderá representar melhorias significativas na qualidade dos sistemas de TA inglês-português atuais. Entre os sistemas que devem se beneficiar com esse módulo está o EPT-Web, ainda em desenvolvimento, para a TA de textos jornalísticos do inglês para o português do Brasil. Apesar de o foco inicial ser a TA, esse módulo poderá também ser empregado, posteriormente, em diferentes aplicações envolvendo a comunicação multilíngüe.

Referências Bibliográficas

- Brown, P., Della Pietra, S., Della Pietra, V. and Mercer, R. (1991) "Word sense disambiguation using statistical methods", In: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, p. 264-270.
- Burnard, L. (2000) "Reference Guide for the British National Corpus (World Edition)", Oxford University Press.
- Dorr, B. J. and Katsova, M. (1998) "Lexical Selection for Cross-Language Applications: Combining LCS with WordNet", In: Proceedings of AMTA'1998, Langhorne, p. 438-447.
- Fossey, M.F., Pedrolongo, T., Martins, R. T. and Nunes, M. G. V. (2004) "Análise comparativa de tradutores automáticos inglês-português", Série de Relatórios do NILC, NILC-TR-04-04, São Carlos, Março, 18p.
- Leffa, V. J. (1998) "Textual constraints in L2 lexical disambiguation", System, Great Britain, 26(2), p. 183-194.
- Lee, H. (2002) "Classification Approach to Word Selection in Machine Translation", In: Proceedings of AMTA'2002, Springer-Verlag, Berlin, p. 114-123.
- Montoyo, A., Romero, R., Vazquez, S., Calle, M. and Soler, S. (2002) "The Role of WSD for Multilingual Natural Language Applications", In: Proceedings of TSD'2002, Czech Republic, p. 41-48.
- Oliveira Jr., O.N., Marchi, A.R., Martins, M.S. and Martins, R.T. (2000) "A Critical Analysis of the Performance of English-Portuguese-English MT Systems", In: Anais do V PROPOR, Atibaia, p. 85-92.
- Pedersen, B. S. (1997) "Lexical ambiguity in machine translation: expressing regularities in the polysemy of Danish Motion Verbs". PhD Thesis, Center for Sprogteknologi, Copenhagen, Denmark.
- Specia, L. and Nunes, M.G.V. (2004) "A ambigüidade lexical de sentido na tradução do inglês para o português – um recorte de verbos problemáticos", Série de Relatórios do NILC, NILC-TR-04-01, São Carlos, Março, 30p.
- Zinovjeva, N. (2000) "Learning Sense Disambiguation Rules for Machine Translation". Master's Thesis in Language Engineering. Department of Linguistics, Uppsala University.