Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

Introdução aos Métodos e Paradigmas de Tradução Automática

Lucia Specia
Lucia Helena Machado Rino

NILC-TR-02-04

Março, 2002

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil


DC
Desenvolvimento para o Futuro
Departamento de Computação


ufscar

Índice

Figuras

Resumo[1]

A Tradução Automática (TA) é uma das aplicações mais antigas da computação e tem se mostrado cada vez mais necessária, principalmente no cenário de globalização atual, onde a língua ainda representa uma barreira para a comunicação e compartilhamento de informações. Este relatório apresenta o contexto histórico da TA e descreve os diferentes métodos e principais paradigmas que podem ser utilizados para o desenvolvimento de sistemas dessa categoria.

[1] Apoio: CAPES

1 Introdução

No cenário de globalização atual, a disseminação cada vez maior de informações multilíngües, principalmente em meio eletrônico, como a *web*, evidencia a necessidade de traduções rápidas, eficientes e baratas, para facilitar a comunicação e o compartilhamento de informações. Nesse contexto, é crescente o interesse por sistemas de tradução automática (TA), ou seja, por sistemas que permitam a tradução por computador de textos de uma língua natural para outra.

A tarefa de TA consiste em se partir de um texto-fonte, isto é, um texto escrito em uma língua natural (ou língua fonte – LF), a fim de se produzir uma versão em um texto-alvo, em uma outra língua natural (ou língua alvo – LA). Segundo Nirenburg (1987), encontrar uma forma de manter o significado, no texto-alvo, o mais próximo possível do significado do texto-fonte é o principal problema do projeto e desenvolvimento de sistemas de TA. A própria tradução humana é considerada uma das mais complexas atividades de escrita (Santos, 1998), sendo mesmo classificada como uma arte (Hutchins, 1998): a cada passo, envolve escolhas pessoais entre alternativas não codificadas. Logo, não é meramente uma questão de substituições diretas de conjuntos de símbolos, mas sim uma questão de se fazer escolhas entre valores interdependentes.

As dificuldades encontradas no desenvolvimento de sistemas de TA devem-se principalmente à necessidade de um conhecimento detalhado do texto-fonte e da situação comunicativa. Nesse contexto, o Processamento das Línguas Naturais (ou PLN) se apresenta como proposta de solução, visando satisfazer os requisitos básicos da TA, quais sejam: interpretação do texto-fonte e produção de uma representação de seu significado – com possível consideração de seu contexto ou situação discursiva – e produção do texto-alvo na LA. A evolução da pesquisa e desenvolvimento em TA foi, na verdade, fortemente influenciada pelas inovações do PLN e da lingüística formal. O PLN, como uma subárea da Inteligência Artificial, fornece uma série de técnicas computacionais para a análise e geração automática de textos em língua natural; a lingüística formal permite que a competência lingüística do falante de uma língua seja descrita em termos de um número finito de regras ou de princípios lingüísticos para gerar um número infinito de sentenças nessa língua.

As pesquisas na área de TA surgiram na década de 40 e, desde então, vários sistemas acadêmicos e comerciais vêm sendo desenvolvidos, alcançando diferentes níveis de sucesso. Esses sistemas se baseiam em diferentes métodos, a saber, TA direta, TA por transferência e TA por interlíngua. Além disso, tais sistemas podem utilizar diversos paradigmas, os quais correspondem a diferentes componentes de representação de conhecimento.

Neste relatório, apresentamos a cronologia do desenvolvimento da TA (seção 2), para então identificar seus principais métodos (seção 3) e paradigmas (seção 4). Conclusões e comentários gerais sobre esse trabalho são apresentados na seção 5.

2 A evolução da tradução automática

A TA é provavelmente a aplicação mais antiga do PLN e também a primeira aplicação não numérica proposta na área da computação, na década de 40, impulsionada pelo grande número de informações disponíveis e pela idéia de que o processo computacional seria tão direto quanto a tradução humana (Nirenburg, 1987). Nessa época, os objetivos de pesquisa eram modestos, devido a limitações de hardware e à inexistência de linguagens de programação de alto nível. A sintaxe era um tema relativamente negligenciado na área de

lingüística e a semântica era geralmente ignorada. Devido a essas limitações, os sistemas de TA apresentavam resultados de baixa qualidade, exigindo um grande envolvimento humano na edição prévia e/ou posterior à execução. Nesse contexto, eram comumente considerados apenas subconjuntos de construções de uma língua natural, limitadas de acordo com regras de gramática e de vocabulário, em domínios específicos, como forma de restrição das entradas dos sistemas (Hutchins, 1998).

Com o início da guerra fria a partir de 1946, a TA passou a ser de grande interesse, principalmente para americanos e ingleses, cujo objetivo era obter informações científicas soviéticas, em geral à distância e o mais rapidamente possível. A primeira aplicação de TA nessa época foi uma calculadora científica que realizava traduções palavra por palavra, ignorando questões lingüísticas. Com ela, era possível identificar o conteúdo de um texto por uma lista de palavras-chave traduzidas, por exemplo. Em 1948, tal sistema foi refinado, para tratar desinências russas durante a análise gramatical. Já no início dos anos 50, procurou-se explorar automaticamente o contexto dos termos manipulados pela calculadora, visando solucionar problemas de ambigüidade semântica. No entanto, essa proposta era bastante equivocada: acreditava-se que os circuitos lógicos das calculadoras seriam capazes de resolver os elementos lógicos da linguagem, auxiliados pela determinação da área à qual a informação pertencia (Mateus, 1995). Diante de tal equívoco, diversos trabalhos foram desenvolvidos considerando-se a necessidade de pré-edição dos textos a serem submetidos à tradução automática.

Diretrizes mais claras para a TA foram delineadas em 1952, no congresso promovido pelo Instituto de Tecnologia de Massachusetts: deveriam ser investigadas a freqüência das palavras nos textos a serem traduzidos, as equivalências lingüísticas e outros aspectos técnicos, para só então se proceder à análise sintática e à construção, propriamente dita, dos programas de tradução correspondentes. Surgiram, assim, as abordagens fundamentais no PLN: as dirigidas por modelagem lingüística. Além disso, determinou-se, como objetivo mais próximo, o desenvolvimento de sistemas que realizassem a tradução entre duas línguas naturais em um único sentido. Considerou-se também, segundo Alfaro (1998), a possibilidade de utilizar uma língua intermediária, neutra, para se realizar a tradução, a qual viria a ser chamada posteriormente de interlíngua.

A primeira experiência de TA real, do russo para o inglês, foi realizada em 1954, na Universidade de Georgetown, com um vocabulário reduzido (250 palavras), textos cuidadosamente selecionados e 6 regras de sintaxe. Essa experiência foi considerada satisfatória. Segundo Hutchins (1998), a partir desse resultado, órgãos que patrocinavam projetos de TA passaram a acreditar que poderiam ser desenvolvidos sistemas que produzissem traduções de boa qualidade em poucos anos. Entretanto, tal nível de qualidade ainda era dependente da evolução de hardware, do surgimento ou refinamento das linguagens de programação de alto nível existentes e, principalmente, do desenvolvimento das pesquisas para a análise sintática, sobretudo referentes à exploração de gramáticas formais, dentre as quais o grande marco, na época, foi a gramática normativa de Chomsky (1957).

A partir de então, as pesquisas em TA passaram a considerar como objetivo o desenvolvimento de sistemas completamente automatizados produzindo traduções de alta qualidade em domínios amplos. A ênfase nas pesquisas tornou-se a busca por teorias e métodos que permitissem alcançar tais objetivos.

No final dos anos 50, além dos americanos, outros países europeus começaram a explorar e investir na TA. Buscava-se ainda transformar os estudos lingüísticos em uma ciência exata, empregando-se métodos matemáticos. No entanto, os primeiros projetos de TA resultantes desses investimentos não alcançaram suas ambiciosas metas. O progresso foi

muito mais lento do que se esperava, devido à complexidade de tratamento computacional dos aspectos formais, teóricos, da lingüística e aos aspectos da própria TA. A lingüística formal não conseguia explicar, por exemplo, os problemas estruturais, funcionais e práticos da TA. Como resultado, houve um descrédito generalizado na TA, culminando com um relatório do ALPAC (*Automatic Language Processing Advisory Committee*) – comitê composto pelos patrocinadores americanos – em 1966, declarando que a TA havia falhado em atingir suas metas, uma vez que não existia nenhum sistema completamente automático capaz de produzir traduções de boa qualidade. Esse relatório também era fortemente negativo com relação às chances futuras de sucesso da TA, o que provocou um corte radical de verbas governamentais norte-americanas. Alguns poucos projetos foram mantidos, agrupados, sobretudo, em três classes: ferramentas computacionais para auxílio à tradução humana, sistemas de TA envolvendo a interação humana e pesquisas teóricas sobre melhorias dos métodos de TA (Hutchins, 1998).

Nirenburg (1987) diz que é importante lembrar que os esforços iniciais tiveram grande importância para o estudo das línguas naturais e do seu processamento via computador, pois contribuíram para o desenvolvimento de várias áreas, dentre as quais destacam-se a Lingüística Moderna, a Lingüística Computacional e a própria Inteligência Artificial. Após o período "negro" do PLN, alguns fatos inovadores reativaram o interesse pela TA no início da década de 80: foi criada a Comunidade Econômica Européia; houve uma explosão da informatização, com grandes avanços de técnicas de computação e da inteligência artificial; as pesquisas e o desenvolvimento de novas teorias no âmbito da lingüística formal (em especial, as teorias de Chomsky) possibilitaram o aprofundamento das investigações no campo da semântica e o processamento automático de várias línguas naturais com base em gramáticas de análise e de geração. Além disso, a TA se enquadrou em um contexto mais realista, no qual aceitava-se que, mesmo imperfeita, ela poderia ser muito útil.

As reflexões sobre as reais possibilidades da TA originaram novas metas e interesses, caracterizando sistemas de diversas naturezas. Sistemas de recuperação de informação, por exemplo, em um ambiente de TA, podem ser úteis mesmo que suas traduções não sejam muito boas. Basta que permitam a compreensão das idéias principais do texto sob exploração (Slocum, 1985). Com foco em situações e objetivos específicos, a TA passou a receber apoio governamental maciço, principalmente na Europa e Japão, no final da década de 80. Acreditava-se em idéias como a de Slocum (1985), de que um sistema de TA de boa qualidade é aquele que gera um texto que permite uma revisão sem grandes problemas e cuja operação completa (incluindo essa revisão) oferece uma boa relação custo-benefício. Como conseqüência, mantiveram-se as áreas de investigação e desenvolvimento de aplicativos computacionais de auxílio à tradução, além de programas de TA prevendo a intervenção humana. Alguns produtos passaram a ser desenvolvidos, como o *Systran* (http://www.systransoft.com) e o *Eurotra* (http://www.ccl.kuleuven.ac.be/about/EUROTRA.html), sistemas americano e europeu, respectivamente, em constante desenvolvimento.

A partir do final dos anos 80, vários fatores contribuíram para a definição de um novo cenário de TA (Hutchins, 1998): 1) diversos sistemas de tradução comerciais foram disponibilizados no mercado, para amplo uso por tradutores humanos profissionais ou por usuários comuns; 2) cresceu significativamente a aquisição de computadores de uso pessoal, com a perspectiva do aumento de ferramentas de comunicação dedicadas; 3) teve início o desenvolvimento de sistemas particulares, de domínio específico; 4) o crescimento das redes de telecomunicação envolvendo muitas línguas conduziu à demanda de dispositivos de tradução, em tempo real, de grandes volumes de dados eletrônicos; 5) a grande

disponibilidade de bancos de dados e recursos de informações em muitas línguas diferentes levou à necessidade de dispositivos de busca e acesso que incorporassem módulos de tradução.

Nesse cenário, o problema computacional da TA foi praticamente superado: diversos serviços de TA são oferecidos na Internet; grupos de pesquisa em TA permitem que o público realize testes *on-line* dos programas em desenvolvimento; os dicionários e as gramáticas necessários para o funcionamento de sistemas de TA podem ser ampliados pelos próprios usuários; a velocidade e a eficiência das consultas aos bancos de dados são cada vez maiores. No entanto, restam importantes questões de cunho lingüístico a resolver (semântico e pragmático-discursivo, principalmente), tais como ambigüidades, referências anafóricas, etc. Como conseqüência, o desenvolvimento de sistemas completamente automatizados, que consideram questões lingüísticas e extralingüísticas de forma profunda, principalmente em domínios abertos ou línguas naturais irrestritas, após mais de 50 anos de pesquisa, ainda é um desafio para a área de TA.

De fato, segundo Kay (1994), ainda hoje alcançam resultados mais práticos e significativos os sistemas de TA desenvolvidos em contextos limitados, como é o caso do *Taum-Meteo* (Isabelle, 1987), utilizado para produzir boletins meteorológicos bilíngües, do inglês para o francês, cuja linguagem é altamente estilizada, regular e específica. Porém, sistemas baseados em sublínguas não constituem interesse para a tradução entre falantes de duas línguas naturais, por serem altamente restritos pela comunidade de uso.

Em domínios abertos, por outro lado, geralmente os textos traduzidos são compreensíveis, mas nem sempre gramaticais e raramente fluentes, implicando a necessidade de revisão humana na fase de pós-processamento.

Alguns sistemas de TA servem de auxiliares para tradutores humanos, no sentido de que realizam uma pré-tradução do texto, a ser editada/refinada pelos tradutores humanos, a exemplo dos tradutores *Trados Workbench* (http://www.trados.com/), *IBM Translation Manager* (http://www-4.ibm.com/software/ad/translat/) e *Déjavu* (http://www.atril.com). Outros, ainda, consideram a pré-edição do documento original, de modo a apresentá-lo em uma linguagem mais simples, como a usada pela Xerox no *Systran* (http://www.systransoft.com), criado inicialmente para traduzir seus manuais técnicos em várias línguas.

Sistemas de TA que consideram alguma forma de edição humana, seja ela feita previamente, durante a tradução ou posteriormente, são chamados de *Human-Aided Machine Translation* ou, simplesmente, HAMT. Quando servem de auxílio à tradução humana, são chamados *Machine-Aided Human Translation* ou, simplesmente, MAHT. Esses últimos incluem ferramentas de acesso a dicionários e enciclopédias, recursos de processamento de textos, verificação ortográfica e gramatical, entre outras (Boitet, 1994).

Atualmente, a *web* certamente é responsável pelo novo incentivo à TA. Com a popularização da Internet, cresceram consideravelmente a oferta e a procura de programas de TA. Diversos sistemas são capazes de traduzir páginas da Internet *on-line*, mensagens de correio eletrônico ou conversas via programas de *chat*.

É nesse contexto de apelo à necessidade de tradução de textos disponíveis de forma eletrônica que se fundamenta o projeto e desenvolvimento de parte significativa dos sistemas de TA atuais, com base em diferentes métodos de tradução. Consideram-se três métodos, descritos a seguir: tradução direta, por transferência e por interlíngua.

3 Métodos de TA

De acordo com Dorr et al. (2000), dois tipos de informação podem ser utilizados para classificar um sistema de TA: seu método e seu paradigma. Os **métodos** referem-se ao projeto de processamento, ou seja, à organização global do processamento e de seus vários módulos, enquanto os **paradigmas** (que são descritos na seção 4) referem-se aos componentes de representação de conhecimento que auxiliam o projeto de processamento global.

Há três diferentes métodos de TA: **TA direta**, **TA por transferência** e **TA por interlíngua**. Esses métodos podem ser agrupados em duas categorias: a **tradução direta** e a **tradução indireta**, esta incluindo os dois últimos métodos.

A Figura 1 (Dorr et al., 2000, p. 13) delineia os diversos níveis de profundidade do conhecimento a ser manipulado por cada um dos métodos. É importante notar que o mesmo conhecimento lingüístico pode ser utilizado por diferentes métodos (por exemplo, o conhecimento semântico pode ser utilizado tanto no método por interlíngua quanto no método por transferência).

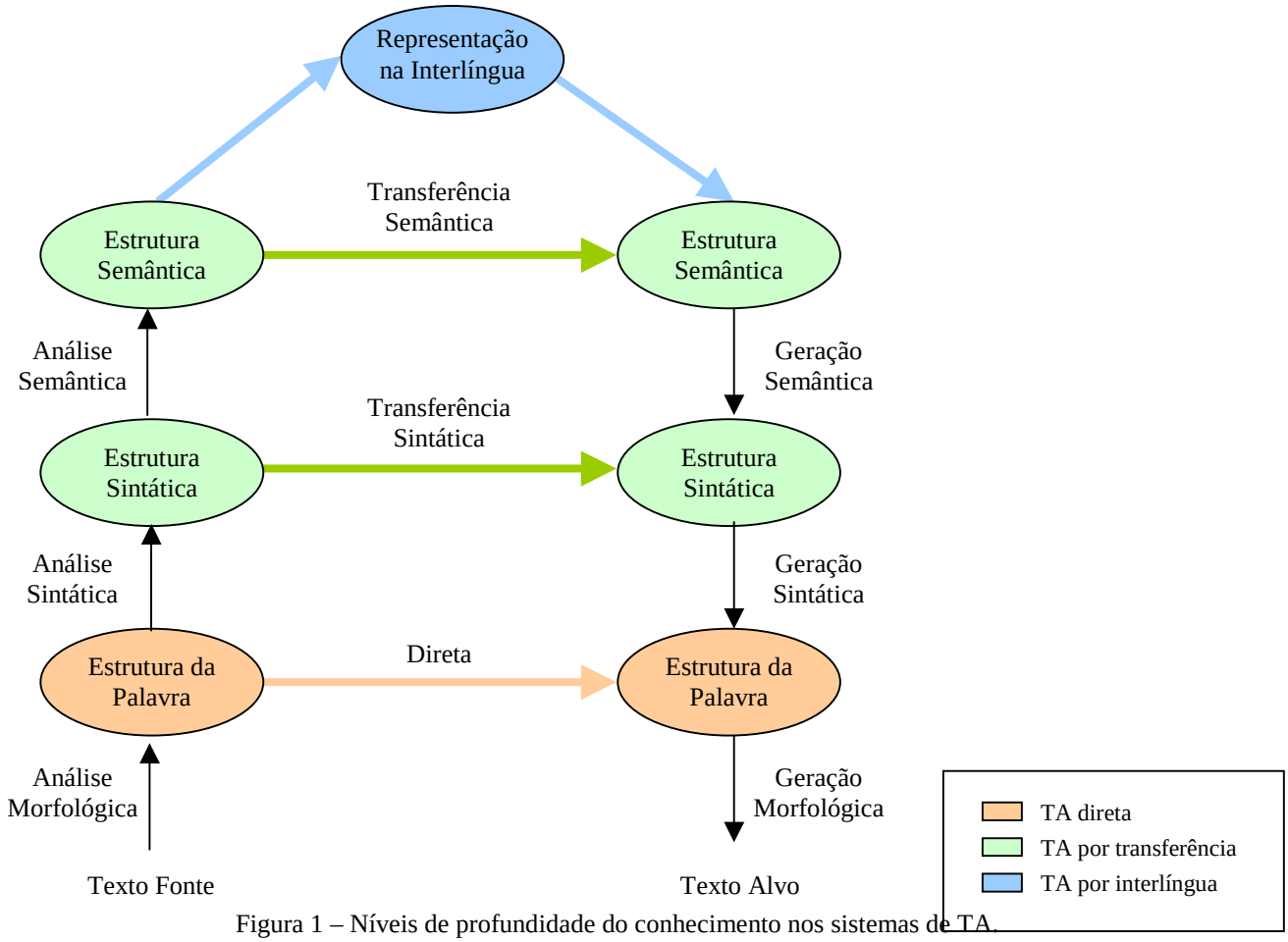

Figura 1 – Níveis de profundidade do conhecimento nos sistemas de TA.

A seguir, são descritas as principais características de cada método.

3.1 Método direto

A TA direta transforma as sentenças da LF em sentenças da LA sem utilizar representações intermediárias, procurando realizar o mínimo de processamento lingüístico possível. Esse processamento pode variar, incluindo a simples substituição das palavras de uma sentença-fonte por sua(s) correspondente(s) na LA (tradução palavra-por-palavra) ou a realização de tarefas mais complexas, como a reordenação das palavras na sentença-alvo e a inclusão de preposições.

Geralmente, o processo de tradução compreende a análise sintática simplificada, a substituição das palavras-fonte por suas equivalentes na LA, utilizando, para tanto, um dicionário bilíngüe, e a reordenação das palavras de acordo com as regras da LA, a partir de informações sintáticas de ambas as línguas (Arnold et al., 1993). A Figura 2 ilustra esse processo.

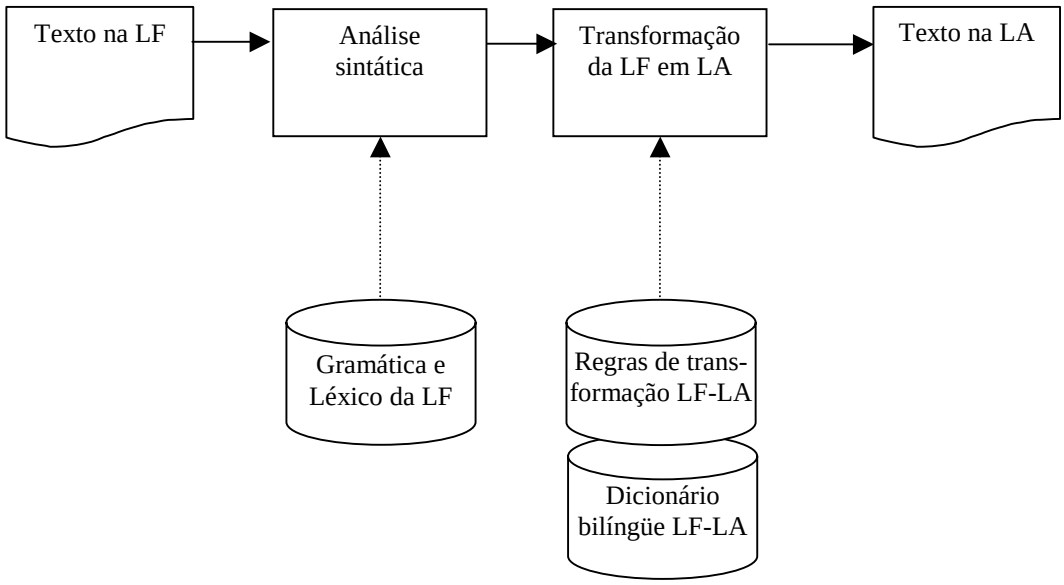

Figura 2 – TA pelo método direto.

Sistemas de TA direta são comumente construídos para um único par de línguas e a definição do número de estágios depende, na verdade, além do nível de qualidade pretendido, da proximidade das línguas envolvidas. Ward (1999) cita um exemplo de um sistema de tradução do japonês para o inglês consistindo de seis estágios:

1) Análise lexical e morfológica, na qual a sentença é separada em palavras;
2) Transferência lexical das palavras na sentença, por meio do dicionário bilíngüe;
3) Adequações relacionadas às preposições (a resolução preposicional do japonês é significativamente distinta da resolução para o inglês, merecendo um módulo isolado);
4) Transformações das estruturas sentenciais. Por exemplo, a estrutura dos constituintes da forma SOV (Sujeito Objeto Verbo), utilizada na língua japonesa, é transformada na estrutura SVO (Sujeito Verbo Objeto), utilizada na língua inglesa;
5) Tarefas diversas, como a inserção de artigos;
6) Geração morfológica, na qual as palavras em inglês são flexionadas.

No que se refere aos resultados esperados por sistemas de TA direta, pode-se identificar as seguintes características (Arnold et al., 1993):

1) O sistema pode ser considerado robusto, pois sempre apresenta algum resultado, mesmo que seja ruim, por conta de problemas como a não tradução de algumas palavras (por não existirem no dicionário), ou a geração de construções gramaticais desconhecidas (por não existirem regras de transformação adequadas).

2) Não há como garantir que a sentença traduzida seja realmente uma sentença gramatical na LA – o resultado pode ser um inelegível emaranhado de palavras.

3) A qualidade do resultado dos sistemas que fazem somente a correspondência direta entre palavras tende a ser muito baixa, o que justifica a exploração de sistemas de tradução direta mais complexos.

Segundo Dorr et al. (2000), de um modo geral, as traduções são pobres; no entanto, se limitadas a domínios restritos e a textos simples, podem ser bastante úteis, principalmente para especialistas naquele domínio.

Quanto à utilização desse método, grande parte dos sistemas de TA comerciais, principalmente os mais antigos, foi desenvolvida com base nele, a exemplo do *Systran* (http://www.systransoft.com), cujo processo de tradução consiste basicamente de buscas em um dicionário, palavra a palavra.

3.2 Método indireto

Nos sistemas de tradução indireta, a análise da LF e a geração da LA constituem processos independentes, cada qual tratando somente dos problemas da língua envolvida. Diferentemente do método direto, esses sistemas se baseiam na idéia de que a TA de alta qualidade requer conhecimento lingüístico (e eventualmente extralingüístico) de ambas as línguas, assim como das diferenças entre elas. Esse conhecimento é representado por linguagens intermediárias entre as línguas fonte e alvo. Assim, as sentenças fonte são primeiramente transformadas numa representação na linguagem intermediária e, a partir dela, são geradas as sentenças alvo.

Existem dois métodos de TA indireta: por transferência e por interlíngua. Dependendo do método, a representação intermediária pode ser única, independente de língua (tradução por interlíngua), ou dependente de língua (tradução por transferência). Neste último caso, são necessárias duas linguagens de representação intermediária: uma para a LF e outra para a LA.

Para codificar o conhecimento das línguas fonte e alvo e o conhecimento das relações entre elas, basicamente são necessários os seguintes componentes:

1) gramáticas e léxicos substanciais de ambas as línguas, os quais são utilizados tanto na análise das sentenças fonte, quanto na geração das sentenças alvo;
2) dicionários bilíngües para as regras de substituição de palavras;
3) no caso da tradução por transferência, uma gramática comparativa, ou seja, um conjunto de regras de transformação para relacionar a representação intermediária da LF com a representação intermediária da LA;
4) no caso de tradução por interlíngua, um conjunto de regras de transformação para relacionar a interlíngua com as línguas fonte e alvo.

Dependendo do método de TA indireta e também do nível de profundidade da análise realizada, outros componentes podem ser necessários, conforme descrito a seguir.

3.2.1 TA por transferência

Na TA por transferência, a tradução consiste nos seguintes passos (Figura 3): 1) alteração da estrutura e palavras da sentença de entrada resultando em uma representação intermediária da LF (fase de análise); 2) transformação dessa representação em uma estrutura intermediária da LA (fase de transferência); e 3) geração da sentença na LA (fase de geração), a partir dessa estrutura.

A fase de análise pode envolver processos complexos como as análises semântica e pragmática, mas, em geral, são mais comuns sistemas que se limitam à análise sintática, gerando como representação intermediária uma estrutura de árvore. Nesse caso, a fase de transferência converte essa estrutura da LF em uma estrutura de árvore da LA, por meio de regras de mapeamento entre as duas línguas naturais, que indicam as correspondências lexicais e sintáticas entre tais estruturas. Para tanto, é necessário representar o conhecimento contrastivo (i.e., comparativo) das duas línguas, o qual envolve a especificação de suas diferenças normativas e lexicais. A fase de geração transforma a estrutura de árvore da LA na sentença final, propriamente dita, utilizando a gramática e o léxico da LA.

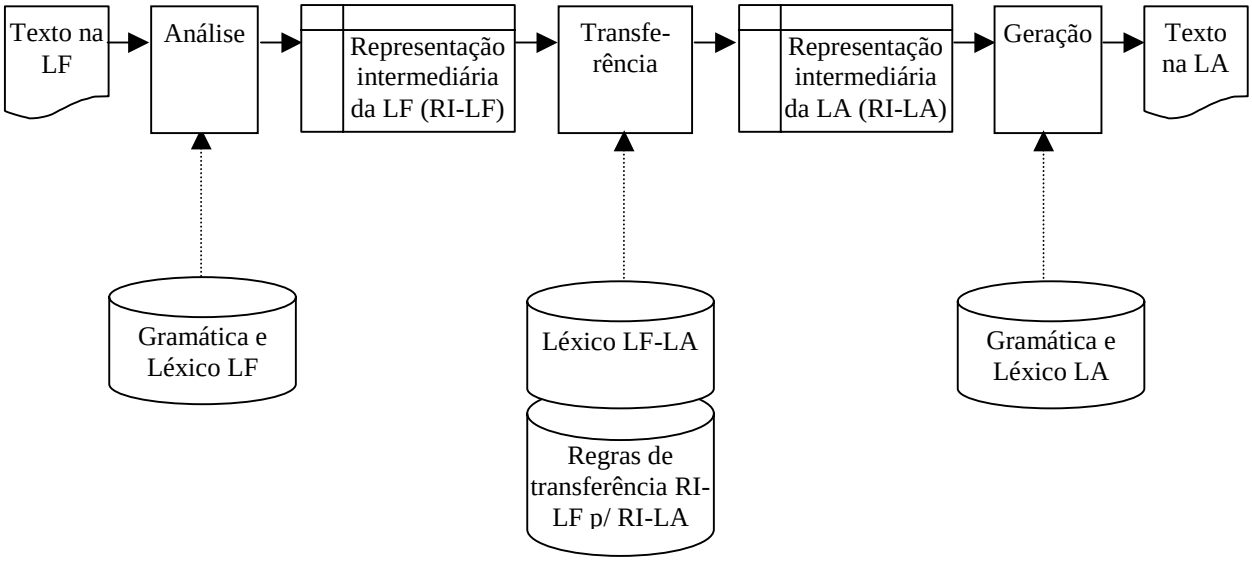

Figura 3 – TA pelo método indireto por transferência

Da mesma forma que na tradução direta, a profundidade do módulo de análise depende da proximidade das línguas: quanto mais próximas forem as línguas, mais superficial pode ser a análise. Porém, para traduções de alta qualidade, mesmo quando as línguas são próximas, análises mais profundas que a sintática são necessárias. Assim, diversos sistemas incorporam aos seus processos também informações semânticas, que podem ser representadas, por exemplo, por *frames*, isto é, coleções de atributos e valores. Nesse caso, um módulo de análise semântica é responsável por preencher os atributos semânticos de componentes sentenciais, por exemplo, a partir de uma árvore sintática; o módulo de transferência mapeia, então, esse *frame* em um outro *frame* da LA, o qual é convertido para a sentença na LA.

Análises ainda mais profundas, baseadas em informações de contexto (pragmático-discursivas) dificilmente são realizadas em sistemas de TA por transferência, dada sua grande complexidade (discutida na seção 3.3). Na verdade, até mesmo as informações de ordem semântica são geralmente incorporadas somente para resolver problemas limitados.

Segundo Dorr et al. (2000), a qualidade global dos sistemas de transferência sintática, é maior que a dos sistemas de tradução direta, mas tende a ser menor que a dos sistemas que empregam uma análise mais profunda do texto fonte, como aqueles que utilizam o método por interlíngua.

Alguns exemplos de abordagens de TA por transferência são o projeto *Eurotra* (http://www.ccl.kuleuven.ac.be/about/EUROTRA.html), cujo objetivo é a criação de um ambiente multilingüe para todas as línguas da Comunidade Européia; e o *Vermobil* (Wahlster, 1993), patrocinado pelo governo da Alemanha, que reconhece textos falados em alemão e os traduz para textos falados em inglês.

3.2.2 TA por interlíngua

Devido à dificuldade de se estabelecer regras de transferência e recursos lingüísticos comparativos (como gramáticas) efetivos, necessários aos sistemas desenvolvidos sob o método de TA por transferência, e também à complexidade inerente a esses sistemas (seção 3.3.), houve o interesse pela definição de um nível de análise tão profundo a ponto de permitir descartar os componentes contrastivos entre as línguas em foco, presentes na tradução por transferência. O objetivo era fazer com que a saída da análise da LF correspondesse diretamente à entrada do componente de geração na LA. Representações nesse nível deveriam capturar, assim, o significado a ser transmitido, independentemente da língua natural em questão. Esta é justamente a função da **interlíngua**: permitir extrair a representação do significado da sentença fonte para, a partir dela, gerar a sentença na LA.

Nesse cenário de TA por interlíngua, o processo de tradução é feito de acordo com as seguintes etapas (Figura 4): 1) análise completa do texto na LF, extraindo seu significado e representando-o na interlíngua; e 2) geração do texto na LA, partindo da representação interlingual e expressando o mesmo significado. Nesse contexto, o processo de geração do texto na LA caracteriza-se mais como uma paráfrase que como uma tradução, podendo ser perdidos o estilo e o foco do texto original.

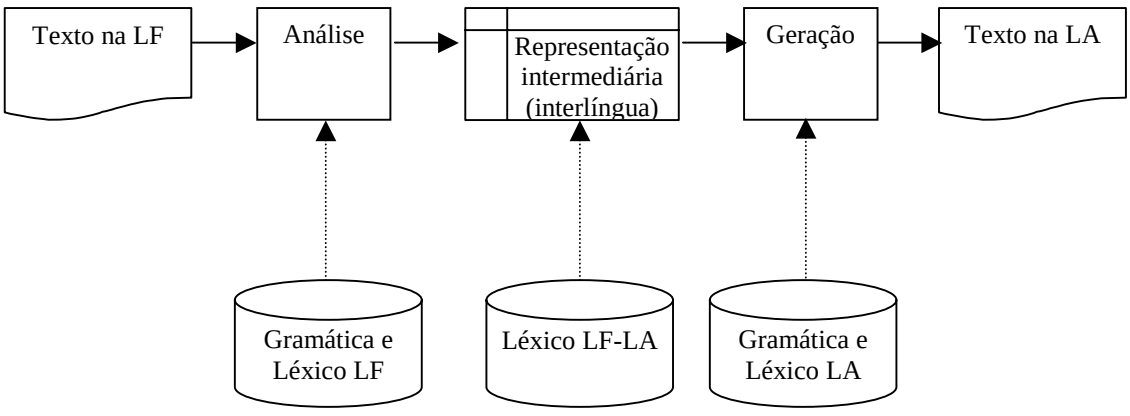

Figura 4 – TA pelo método indireto por interlíngua.

Uma das maiores dificuldades da TA por interlíngua é a própria especificação da interlíngua, a qual deve ser independente de qualquer língua natural, para representar o significado de suas sentenças de modo uniforme e consistente. Os principais problemas são: 1) como escolher o léxico dessa interlíngua, que deve ser composto por conceitos primitivos que permitam expressar o significado de todo o vocabulário das línguas em questão e 2) como

definir a gramática da interlíngua, a qual deve conter regras suficientemente robustas para relacionar todos os possíveis conceitos primitivos.

De acordo com Ward (1999), é preciso realizar análises exaustivas sobre a semântica do domínio, de modo a formalizar os tipos de entidades que existem e o seu relacionamento. Isso pode ser facilitado a partir da definição de uma **ontologia**, isto é, de um modelo de mundo sobre um dado domínio, contendo todos os conceitos (palavras) que podem ser expressos naquele domínio e suas relações (Uschold e Gruninger, 1996).

As principais vantagens da TA por interlíngua são as seguintes:

1) A facilidade com que os sistemas podem ser estendidos, pois novas línguas podem ser adicionadas a um custo relativamente baixo, conforme descrito na seção 3.3; e

2) A possibilidade de incluir níveis de representação mais profundos, como o pragmático-discursivo, resultando em sistemas com potencial maior qualidade que os desenvolvidos sem esses níveis.

Uma vez que sistemas de tradução por interlíngua requerem um conhecimento extensivo, seu desempenho depende, em grande parte, da coleção e representação eficiente de grandes quantidades de conhecimento sobre o domínio em questão. Isso deve ser feito por meio de teorias ou linguagens de representação (semânticas e/ou pragmático-discursivas), como, por exemplo, a Teoria Dependência Conceitual (Schank, 1975), a Teoria de Casos (Fillmore, 1968), a Semântica Conceitual (Jackendoff, 1990) e a UNL (UNL, 2001).

Em vista das vantagens oferecidas pela TA por interlíngua, nos últimos anos vários sistemas baseados nesse método vêm sendo desenvolvidos, envolvendo diferentes línguas e objetivos específicos. O foco das pesquisas consiste, geralmente, na especificação de uma interlíngua que seja adequada para todas as línguas envolvidas, além de uma análise conceitual que permita produzir traduções de qualidade aceitável (Dorr et al., 2000). A seguir, são descritos brevemente alguns desses sistemas.

Exemplos de sistemas de TA por interlíngua

- **TRANSLATOR** (Nirenburg et al., 1987): sistema que explora o paradigma KBMT (discutido na seção 4.1), utilizando, além dos módulos de análise e geração dos sistemas de TA por interlíngua, um módulo de *enriquecimento*, o qual dispõe de uma base de conhecimento (um dicionário e uma gramática da interlíngua) para, a partir do texto na interlíngua gerado pelo analisador, realizar algumas inferências de modo a acrescentar a esse texto informações sobre o domínio em questão. A estrutura enriquecida é então repassada ao gerador da LA. O TRANSLATOR possui duas linguagens de representação do conhecimento: uma para descrever os conceitos da interlíngua no dicionário (DRL) e uma para descrever a sintaxe da interlíngua na gramática (GRL).

- **KBMT-89** (Nirenburg e Goodman, 1991): sistema para traduções entre inglês e japonês, baseado em conhecimento (KBMT), tendo como domínio manuais de manutenção e instalação de computadores pessoais. A representação da interlíngua é feita por *frames*. O sistema se baseia no conhecimento expresso por uma ontologia muito bem definida, criada a partir do formalismo *Ontos* (Carlson e Nirenburg, 1992).

- **ULTRA** – *Universal Language TRAnslator* (Farwell e Wilks, 1991): sistema de TA multilingüe que realiza traduções de palavras individuais ou de sentenças entre as línguas chinesa, inglesa, alemã, espanhola e japonesa. Contempla sentenças declarativas, interrogativas e imperativas e construções conjuntivas, além de tratar de algumas desambigüizações de sentido e dependências contextuais (anafóricas e elípticas, por exemplo).

- **KANT** (Mitamura et al., 1991; Nyberg e Mitamura, 1992): sistema de TA multilingüe, baseado no paradigma KBMT, que traduz documentos técnicos do inglês para o japonês, francês e alemão, com qualidade elevada. O domínio das traduções é específico (manuais técnicos) e a entrada para o sistema é na forma de textos em uma linguagem simplificada (isto é, os textos são pré-processados). As principais características da arquitetura do sistema são a aquisição de conhecimento semi-automatizada e a interpretação semântica profunda.

- **UNITRAN** (Dorr, 1992): sistema baseado na semântica lexical, desenvolvido para traduções bidirecionais entre espanhol, inglês e alemão. A representação interlingual adotada é a LCS (*Lexical Conceptual Structure*), uma extensão da estrutura conceitual de Jackendoff (1990), escolhida principalmente por prover um mapeamento sistemático entre a interlíngua e as estruturas sintáticas.

- **UNL** – *Universal Networking Language* (Uchida et al., 1999; UNL, 2001): projeto que contempla a TA multilingüe, incluindo atualmente 14 línguas, envolvendo o trabalho cooperativo de diversos países. Para esse projeto, a UNU (Universidade das Nações Unidas) criou e disponibilizou para todos os grupos de projeto e desenvolvimento UNL os seguintes recursos: 1) a especificação da interlíngua UNL; 2) codificadores (da LF para a IL) e decodificadores (da IL para a LA) genéricos a serem customizados para cada língua natural em foco; e 3) uma base de conhecimento para representar informações universais a respeito de todos os conceitos do repositório da UNL, inclusive informações ontológicas, na forma de uma hierarquia conceitual.

3.3 A complexidade dos métodos de TA

A comparação da complexidade dos três métodos de TA apresentada aqui se dá em função do número de módulos necessários ao processo de tradução. Considera-se que os sistemas de TA podem trabalhar com tradução **unidirecional**, isto é, para cada par de línguas, uma língua é fonte ou alvo, mas não as duas coisas; ou **bidirecional**, isto é, para cada par de línguas L1 e L2, a tradução pode ocorrer tanto de L1 para L2, quanto de L2 para L1. Em ambos os casos, assume-se que tais módulos realizam análise/transferência/geração em apenas um sentido, por exemplo, um módulo para transferência entre representações intermediárias do português para o inglês não realiza a transferência entre representações intermediárias do inglês para o português.

No método direto, considerando apenas o módulo de transformação da LF em LA, para cada par de línguas é necessário um módulo de transferência, se a tradução for unidirecional. Assim, para *n* línguas, são necessários *n-1* módulos. Se a tradução for bidirecional, são necessários dois módulos de transferência, totalizando *n*(n-1)* módulos.

O método por transferência, para *n* línguas, na tradução unidirecional, envolve *n-1* módulos de transferência, além de *1* módulo de análise e *n-1* módulos de geração, totalizando *2n-1* módulos. Já para a tradução bidirecional são necessários *n*(n-1)* módulos de transferência, *n* módulos de análise, e *n* módulos de geração, somando *n*(n+1)* módulos.

No método por interlíngua, a quantidade de módulos necessários é proporcional ao número de línguas que o sistema manipula, e não ao quadrado desse número, como no método por transferência. Para cada nova língua, são criados apenas os módulos de análise e/ou geração, ou seja, as regras de mapeamento da nova língua para a interlíngua e/ou vice-versa. Assim, para *n* línguas, são necessários *n* módulos para a tradução unidirecional e *2n*, para a tradução bidirecional.

De modo sintetizado, a Tabela 1 apresenta as características de complexidade dos três métodos de tradução para *n* línguas, considerando sistemas de tradução uni ou bidirecional.

Método	**Nº de módulos (para tradução unidirecional)**	**Nº de módulos (para tradução bidirecional)**
Direto	n-1	n*(n-1)
Transferência	n^2	n*(n+1)
Interlíngua	n	2n

Tabela 1 – Complexidade dos métodos de TA

4 Paradigmas de TA

Sistemas desenvolvidos de acordo com os diferentes métodos descritos podem se basear em conhecimento profundo, fundamental ou lingüístico – **paradigma fundamental** – ou em conhecimento superficial ou empírico – **paradigma empírico** (Arnold et al., 1993). Esta seção descreve, brevemente, em 4.1, os modelos de representação do conhecimento fundamental e em 4.2, algumas abordagens empíricas atuais.

As combinações possíveis entre métodos e paradigmas são várias. O uso de um paradigma nem sempre exclui o uso de outros, pelo contrário, muitos sistemas de TA são baseados em abordagens híbridas, conforme consta na seção 4.3, as quais podem incluir também combinações entre diferentes métodos, cada método sendo responsável pelo tratamento de determinados aspectos da tradução (sistemas *multi-engine*).

4.1 Paradigmas fundamentais

Os modelos de TA fundamentais são aqueles que empregam teorias lingüísticas bem definidas, utilizando restrições sintáticas, lexicais ou semânticas, sobre as línguas naturais envolvidas. A seguir, são descritos alguns dos diferentes tipos de conhecimento que caracterizam esses modelos.

4.1.1 TA baseada em regras

Sistemas de TA baseados em regras (*Rule-Based Machine Translation*, ou RBMT) são caracterizados por representar o conhecimento por meio de regras de diferentes níveis lingüísticos, para a tradução entre as línguas fonte e alvo. Por exemplo, para a transferência

lexical, as características e restrições de itens lexicais individuais são codificadas num mecanismo de controle, por meio de regras, e não no léxico.

Rosseta (1994) descreve um exemplo de sistema RBMT interlingual, dividindo as regras de tradução em duas categorias: 1) regras que fazem o mapeamento de árvores sintáticas em estruturas de significado; e 2) regras que fazem o mapeamento de itens lexicais em árvores sintáticas.

4.1.2 TA baseada em conhecimento

O paradigma baseado em conhecimento (*Knowledge-Based Machine Translation*, ou KBMT) define sistemas baseados em regras que utilizam conhecimento profundo, lingüístico ou extralingüístico, de um domínio, permitindo que o sistema possa tecer inferências sobre os conceitos manipulados. Segundo Kay (1994), a maior justificativa para utilização de sistemas KBMT é que a tradução depende fortemente de informações e características extralingüísticas, de senso comum e de conhecimento do mundo. A representação do conhecimento pode envolver o desenvolvimento de ontologias e modelos de domínio. Alguns exemplos de sistemas dessa categoria são o *Translator* (Nirenburg et al., 1987), o KBMT-89 (Nirenburg e Goodman, 1991), o KANT (Mitamura et al., 1991; Nyberg e Mitamura, 1992) e o Projeto UNL (Uchida, 1999; UNL, 2001), os quais foram brevemente apresentados na seção 3.2.2.

4.1.3 TA baseada em léxico

Sistemas baseados em léxico (*Lexicon-Based Machine Translation*, ou LBMT) são aqueles que fornecem regras para relacionar as entradas lexicais de uma língua às entradas lexicais de outra língua. Um exemplo de sistema dessa categoria é o LTAG (Abeillé et al., 1990)[2], para traduções do inglês para o francês e vice-versa. O LTAG é um sistema de transferência que utiliza TAGs – *Tree Adjoining Grammars* (Joshi, 1987) para mapear derivações TAG superficiais de uma língua para outra. O mapeamento é realizado por meio de um léxico bilíngüe que associa diretamente árvores fonte e alvo por meio de ligações entre itens lexicais e seus argumentos. De modo simplificado, cada entrada nesse léxico bilíngüe contém regras para o mapeamento entre a sentença na LF e a sentença na LA.

4.1.4 TA baseada em restrições

O paradigma baseado em restrições (*Constraint-Based Machine Translation*, ou CBMT) permite definir restrições em vários níveis de descrição lingüística, por exemplo, para os itens lexicais. Entre as abordagens que utilizam esse paradigma, estão os sistemas de TA que combinam a LFG (Kaplan e Bresnan, 1982), com restrições sobre os itens lexicais, como o LFG-MT (Kaplan et al., 1989)[3]. Nesse sistema, as operações de mapeamento requeridas na

[2] Abeillé, A.; Schabes, Y.; Joshi, A.K. (1990). Using Lexicalized Tags for Machine Translation. In *Proceedings of Thirteenth International Conference on Computational Linguistics (COLING – 90),* pp. 1-6. Helsinki, Finland. Apud (Dorr et al., 2000), p. 23.

[3] Kaplan, R.; Netter, K.; Wedeking, A.Z.J. (1989). Translation by Structural Correspondence. In *Proceedings of Thirteenth International Conference on Computational Linguistics (COLING – 90).* Helsinki, Finland. Apud (Dorr et al., 2000), p. 19.

transferência são executadas por equações de transferência baseadas em restrições que relacionam estruturas-f (estruturas funcionais da LFG) fonte e alvo.

4.1.5 TA baseada em princípios

Sistemas PBMT (*Principle-Based Machine Translation*) são uma alternativa aos sistemas RBMT, nos quais as regras são substituídas por um pequeno conjunto de princípios que envolvem fenômenos morfológicos, gramaticais e lexicais, de um modo geral. Um exemplo de construção derivada de princípios gerais é a construção da voz passiva, conforme descrito por Berwick (1991)[4]. Como não existe uma única regra de mapeamento entre duas línguas naturais para a voz passiva, é comum utilizar-se um conjunto de princípios que definem as operações morfológicas e sintáticas necessárias.

O *Princitran* é um sistema PBMT (Dorr et al., 1995)[5], baseado nos princípios sintáticos da Teoria da Regência e Ligação (*Government-Binding*, ou GB – Chomsky, 1981) e nos princípios semântico-lexicais da LCS (Dorr, 1993). Nesse sistema, a construção de estruturas é adiada até que as descrições satisfaçam os princípios lingüísticos.

O paradigma PBMT é complementar às abordagens KBMT e EBMT, no sentido de que ele provê uma cobertura ampla para muitos fenômenos lingüísticos, mas lhe falta conhecimento mais profundo sobre o domínio de tradução.

4.1.6 TA *shake and bake*

O S&BMT (*Shake & Bake Machine Translation*) (Beaven, 1992)[6] é um dos paradigmas de tradução mais recentes. Ele utiliza regras de transferência como mecanismo para realizar a tradução, mas enquanto o mapeamento entre itens lexicais é realizado por meio de regras de transferência padrão, o algoritmo para combinar esses itens para formar uma sentença na LA não é convencional (Dorr et al., 2000).

As regras de transferência são definidas com base em entradas lexicais bilíngües, que relacionam itens monolingües. Após a análise da sentença da LF, suas palavras são mapeadas em palavras da LA por meio das entradas bilíngües. O algoritmo que combina as palavras na LA tenta ordená-las baseando-se nas restrições sintáticas da LA.

Para construções complexas, como os casos de troca de núcleo, diferentemente da abordagem por transferência simples, o paradigma S&BMT é capaz de construir regras de mapeamento não composicionais selecionando as palavras na LA a partir de um léxico bilíngüe e tentando diferentes ordenações para essas palavras (*shake*) que satisfaçam todas as restrições sintáticas, até que a sentença seja produzida (*bake*).

Essas regras formam a base para a transferência entre as entradas lexicais na LF e LA. A idéia central desse paradigma é que, uma vez que os elementos bilíngües identifiquem corretamente os índices das entradas lexicais, um algoritmo S&BMT pode combiná-los. O principal benefício dessa abordagem é que os léxicos bilíngües precisam somente especificar

[4] Berwick, R. C. (1991). Principles of Principle-Based Parsing. In R.C. Berwick, S.P. Abney, and C. Tenny, editors, *Principle-Based Parsing: Computation and Psycholinguistics*, pp. 1-37. Kluwer Academic Publishers. Apud (Dorr et al., 2000), p. 26.

[5] Dorr, B.J.; Lin, D.; Lee, J.; Suh, S. (1995). Efficient Parsing for Korean and English: A Parameterized Message Passing Approach. *Computational Linguistics*, 21(2), pp. 255-236. Apud (Dorr et al., 2000), p. 27.

[6] Beaven, J. Shake and Bake Machine Translation. In *Proceedings of Fourteenth International Conference on Computational Linguistics.* Nantes, France, pp. 603-609. Apud (Dorr et al., 2000), p. 28.

o conhecimento contrastivo entre duas línguas; as gramáticas monolingües usadas para o *parser* e geração se responsabilizam pelo restante (Dorr et al., 2000). A desvantagem dessa abordagem é que a geração é um problema NP-completo, ou seja, não há um algoritmo eficiente para geração de uma estrutura S&BMT.

4.2 Paradigmas empíricos

Os paradigmas empíricos são os que utilizam pouca ou nenhuma teoria lingüística no processo de tradução. Em geral, eles indicam técnicas experimentais para especificar o mecanismo de tradução apropriado ao contexto em foco. Esses paradigmas passaram a ser bastante explorados nos últimos anos devido ao grande avanço de hardware e à disponibilidade crescente de recursos eletrônicos significativos (dicionários, corpora de textos bilíngües e monolíngües, etc.), componentes essenciais para o sucesso da investigação empírica.

4.2.1 TA baseada em estatística

Sistemas baseados em estatística (*Statistical-Based Machine Translation*, ou SBMT) utilizam técnicas estatísticas ou probabilísticas que contemplam as tarefas lingüístico-computacionais em foco na tradução (por exemplo, a desambigüização lexical).

A idéia dessa abordagem é que a tradução seja realizada por meio de dados estatísticos extraídos automaticamente de corpora de textos bilíngües paralelos. Alguns exemplos de dados que podem ser obtidos a partir da análise desses corpora são:

- probabilidade de uma sentença fonte ocorrer no texto-alvo;
- probabilidade de uma palavra fonte ser traduzida como uma, duas ou mais palavras alvo;
- probabilidade de tradução de cada palavra em outra palavra da língua alvo;
- probabilidade da posição de cada palavra na sentença na língua fonte, quando essa posição não é a mesma que a da palavra na sentença alvo.

As probabilidades obtidas são utilizadas para calcular como uma sentença fonte pode ser traduzida em uma sentença alvo. Há diversas formas de realizar esse cálculo, por exemplo, ele pode ser baseado numa variante da Regra de Bayes, que equaciona o problema da tradução como a produção de uma saída que maximize um valor funcional (Pr(A|F)), este representando a importância de se manter a fidelidade ao texto original e a fluência do texto traduzido. Nesse método, a probabilidade de uma sentença alvo (A) ser a tradução de uma dada sentença fonte (F) é proporcional ao produto da probabilidade de que a sentença A seja uma construção legal na LA (fluência) e da probabilidade de que uma sentença na LF seja a tradução da sentença na LA (fidelidade). A seguinte equação expressa essa relação (Dorr et al., 2000):

$$Pr(A|F) = Pr(A) * Pr(F|A)$$

Um exemplo de sistema que utiliza esse modelo de processamento estatístico é o *Candide* (Brown, 1990), de tradução do francês para o inglês. Esse sistema considera que a probabilidade de qualquer palavra na sentença alvo ser parte de uma sentença legal depende das probabilidades de ocorrência das duas palavras anteriores e que a probabilidade de que a

sentença inteira seja uma sentença legal é o produto de ocorrência de todas as triplas de palavras em um corpus de textos em inglês. Já a probabilidade de que uma palavra na LF seja uma tradução de uma dada palavra na LA depende somente da probabilidade da ocorrência da palavra em uma sentença alvo, de acordo com as probabilidades de alinhamento dos pares de sentenças no corpus.

Um dos problemas da abordagem estatística é a necessidade de corpora de textos substanciais e de boa qualidade, o que torna as traduções muito dependentes do domínio do corpus. Outro problema é que a única forma de melhorar a qualidade da tradução é melhorar a exatidão dos modelos probabilísticos da língua alvo e do processo de tradução, o que exige a adição de muitos parâmetros, além dos já requeridos pelos vários modelos disponíveis. Uma alternativa para amenizar ambos os problemas são os sistemas híbridos, descritos na seção 4.3.

Uma descrição mais detalhada sobre o processo de criação de sistemas estatísticos de TA pode ser consultada em Knight (1999); em Borthwick (1997) são apresentadas três diferentes formas de modelar a TA estatística: *N-grams,* Árvore de Decisão e Entropia Máxima.

4.2.2 TA baseada em exemplos

Na abordagem EBMT (*Example-Based Machine Translation*), também chamada de **TA baseada em casos**, em vez de regras de mapeamento entre as línguas, utiliza-se um procedimento que tenta combinar o texto a ser traduzido com exemplos de traduções armazenados. A tradução é, portanto, por analogia com exemplos coletados a partir de traduções já realizadas, os quais são anotados com suas descrições superficiais, em um corpus bilíngüe alinhado.

Basicamente, a idéia é utilizar um algoritmo de unificação para encontrar o exemplo mais próximo da sentença de entrada, a partir do corpus bilíngüe. Esse procedimento resulta num *template* de tradução, o qual pode, então, ser preenchido palavra-por-palavra, de acordo com as palavras da sentença de entrada.

A proximidade de cada exemplo com a sentença de entrada é determinada pela distância semântica entre as suas palavras, a qual pode ser calculada com base na distância entre essas palavras em uma hierarquia de termos e conceitos provida, em geral, por um *thesaurus* ou uma ontologia.

A combinação de frases requer pelo menos uma análise sintática básica das traduções paralelas, além de alguma análise semântica para determinar a proximidade da combinação. Assim, a tradução de sentenças exige também que a estrutura sintática da sentença fonte seja combinada com sentenças no corpus. A maioria dos sistemas EBMT não considera a combinação da sentença inteira, mas sim de algumas de suas partes, como sintagmas nominais ou preposicionais.

A exatidão e a qualidade da tradução dos sistemas que utilizam o paradigma EBMT dependem da existência de um bom conjunto de dados. A grande cobertura de divergências sintáticas e semânticas requerida pode resultar em um conjunto de informações cujo tamanho dificulta o armazenamento e as buscas. A combinação do paradigma EBMT com abordagens lingüísticas (especialmente com sistemas RBMT) permite diminuir o tamanho desse conjunto de informações. Alguns métodos para adicionar conhecimento lingüístico a sistemas EBMT são descritos por Brown (1999).

Uma das vantagens desse paradigma é que a qualidade da tradução pode melhorar de forma incremental à medida que os exemplos tornam-se mais completos, sem a necessidade de atualizar ou melhorar descrições lexicais ou gramaticais. Algumas complicações nesse modelo ocorrem, por exemplo, quando se tem um número diferente de exemplos e cada um combina com uma parte da sentença, mas as partes que eles combinam se sobrepõem (Arnold et al., 1993). Um exemplo de sistema que utiliza o paradigma EBMT é o *Pangloss* (Brown, 1996).

4.2.3 TA baseada em diálogo

Sistemas de TA baseados em diálogo (*Dialogue-Based Machine Translation*, ou DBMT) são voltados para usuários que são os autores do texto a ser traduzido. Esse tipo de sistema provê um mecanismo que estabelece um diálogo sobre a tradução com o usuário, permitindo que este desambigüise o texto de entrada e incorpore detalhes estilísticos para obter uma tradução de melhor qualidade. Sistemas DBMT são similares aos EBMT, no sentido de que uma representação básica do texto de entrada do usuário é construída e, à medida que ela é revisada por meio de diálogos iterativos com o usuário, são feitas tentativas para atualiza-la a partir de informações armazenadas em um banco de dados de traduções. Além da interação com o usuário durante o processo de tradução, como um mecanismo de desambigüização *on-line* guiado pelo usuário, essa interação pode ocorrer antes do texto de entrada ser repassado ao sistema, como uma forma de revisão prévia guiada pelo usuário.

Em alguns sistemas são codificadas informações de contexto, de modo que o sistema possa determinar a provável intenção do usuário. Usando essas informações, o usuário pode ser guiado por uma série de pontos de escolha, os quais permitem a construção de uma representação que é oferecida ao sistema como candidata à tradução.

Assim como os sistemas KBMT e EBMT, sistemas DBMT são mais voltados para domínios bastante restritos. Para domínios mais abrangentes, a quantidade de informações requerida é muito grande, o que dificulta o armazenamento e a busca.

Um exemplo de sistema desenvolvido com base no paradigma DBMT é o *ENtran* (Johnson e Whitelock, 1987), projetado para prover a construção de um texto de entrada restrito que, traduzido, deixa vários fenômenos lingüísticos para serem processados pelo usuário.

4.2.4 TA baseada em redes neurais

A incorporação da tecnologia de redes neurais e abordagens conexionistas na TA (*Neural-Based Machine Translation*, ou NBMT) é uma área de pesquisa relativamente nova. Essa tecnologia tem sido utilizada basicamente nas funções de *parser*, desambigüização lexical e aprendizado de regras de gramática, considerando-se subconjuntos bastante restritos das línguas. A manipulação de grandes vocabulários e gramáticas aumenta demasiadamente o tamanho das redes neurais e dos conjuntos de treinamento e, conseqüentemente, do tempo de treinamento.

Segundo Dorr et al. (2000), apesar das várias pesquisas sobre esse paradigma, nenhum sistema real de TA foi construído baseado somente na tecnologia de redes neurais, por isso, essa é considerada mais uma técnica auxiliar para a TA.

4.3 Paradigmas híbridos

Muitos paradigmas, principalmente os empíricos, apresentam dificuldades para manipular alguns aspectos do processo de TA. Por exemplo, sistemas estatísticos (SBMT) não manipulam dependências conceituais de longa distância, enquanto que sistemas baseados em exemplos (EBMT) dificilmente tratam estruturas sentenciais complexas. Assim, é reconhecida a necessidade de combinar paradigmas de forma a explorar as vantagens de cada um.

Um exemplo de abordagem híbrida comum consiste em se utilizar paradigmas lingüísticos para a análise automática de textos-fonte e paradigmas estatísticos ou baseados em exemplos para resolver as traduções frasais e as interdependências de constituintes.

Paradigmas estatísticos e probabilísticos devem predominar se o objetivo for obter robustez e grande cobertura de dados. Já se o objetivo for tratar de detalhadas nuances da língua, devem predominar os paradigmas fundamentais. A forma exata de como combinar os diferentes módulos em um sistema, no entanto, permanece uma questão em aberto.

Brown e Frederking (1995), por exemplo, propõem o uso de informações estatísticas para melhorar os resultados da TA fundamental, já Och e Weber (1998) propõem o uso de categorias e regras para melhorar a TA estatística, enquanto Al-Onaizan et al. (1999) utilizam a análise de dados estatísticos para adquirir conhecimento lingüístico de forma automática a partir de corpora bilíngües.

Exemplos de abordagens híbridas são: o sistema *Pangloss* (Brown, 1996), que utiliza os paradigmas EBMT e KBMT juntamente com o método por interlíngua; o sistema *Lingstat* (Barnett et al., 1994)[7], que utiliza o método de transferência para a tradução do japonês para o inglês, com o uso de uma gramática livre de contexto probabilística; e o sistema *Vermobil* (Vogel et al., 2000; Vogel et al., 2000b), também de TA por transferência, que apresenta um módulo auxiliar estatístico.

É importante ressaltar que um modelo híbrido pode envolver não somente a combinação de paradigmas de TA, mas também a combinação de métodos de TA (sistemas *multi-engine*).

5 Conclusões e comentários finais

O interesse pela área de TA existe desde a década de 40, porém, tem se intensificado nos últimos anos, em função da grande quantidade de informações disponíveis principalmente em meio eletrônico e da crescente necessidade de comunicação entre pessoas de diferentes línguas.

As pretensões iniciais para esses sistemas, bastante ambiciosas, foram adequadas de acordo com as limitações de hardware/software e, principalmente, as limitações de cunho lingüístico. Hoje, as maiores restrições aos sistemas de TA são impostas praticamente pela falta de solução computacional para diversos problemas lingüísticos.

Os diversos sistemas de TA existentes apresentam diferentes finalidades, domínios e abrangência, variando desde sistemas simples, de tradução de palavras individuais, a sistemas mais complexos, que consideram informações semânticas e contextuais. Geralmente sistemas simples são de finalidade mais geral, em domínios mais abertos, abrangendo um grande número de dados. Porém, apresentam pior desempenho do que sistemas mais complexos, os

[7] Barnnet, F.; Cant, J.; Demedts, T.D.; Gates, B.; Hays, E.; Ito, Y.; Yamron, J. (1994). LINGSTAT: State of the System. In *ARPA Workshop on Machine Translation*. Vienna, Virginia. Apud (Dorr et al., 2000), p. 32.

quais, devido à quantidade de informações de que necessitam, costumam ser limitados a domínios específicos, com objetivos bem definidos e menor abrangência. Essas limitações permitem que esses últimos sistemas obtenham resultados consideravelmente mais satisfatórios que os primeiros, em termos de qualidade das traduções.

Nos últimos anos, foi possível perceber, principalmente em sistemas de grande porte, uma forte preocupação com a consideração de informações sobre o significado dos textos a serem traduzidos, sejam elas de natureza semântica, pragmática, ou de senso comum, de modo a melhorar a qualidade das traduções. Nesse contexto, o método de TA por interlíngua é o que se mostra mais adequado, por permitir a representação do conhecimento de forma abstrata, independente de língua e possibilitar o desenvolvimento de ambientes multilíngües com uma complexidade relativamente baixa, se comparada à de outros métodos.

Pôde-se observar também, no desenvolvimento de sistemas de TA, uma tendência a considerar combinações de diferentes métodos e/ou paradigmas, resultando em abordagens híbridas, para obter traduções de melhor qualidade, como por exemplo, o uso do paradigma estatístico (SBMT) como processamento complementar ao realizado por métodos de transferência.

De um modo geral, apesar da considerável evolução da TA, seus resultados ainda precisam ser bastante aprimorados. Nesse sentido, mantém-se a dependência de teorias lingüísticas bem definidas, para diferentes línguas, e de estudos na Lingüística Computacional para descobrir meios de implementa-las.

Referências bibliográficas

Al-Onaizan, Y. et al.. (1999). *Statistical Machine Translation*. Final Report. In *Johns Hopkins University 1999 Summer Workshop on Language Engineering*, Center for Speech and Language Processing, Baltimore.

Alfaro, C. (1998). *Descobrindo, Compreendendo e Analisando a Tradução Automática*. Monografia de Conclusão de Especialização. PUC, Rio de Janeiro.

Arnold, D.J.; Balkan, L.; Humphreys, R.L.; Meijer, S.; Sadler, L. (1993). *Machine Translation: an Introductory Guide*. Blackwells-NCC, London.

Boitet, C. (1994). (Human-Aided) Machine Translation: A Better Future?, Grenoble.

Borthwick, A. (1997). *Survey Paper on Statistical Language Modeling*. Tech. Report, New York University, New York.

Brown, P.F. (1990). A Statistical Approach to Machine Translation. In *Computational Linguistics*, 16(2).

Brown, R.D. (1996). Example-Based Machine Translation in the Pangloss System. In *Proceedings of the 16th International Conference on Computational Linguistics* - COLING-96, pp. 169-174. Copenhagen, Denmark, August 5-9.

Brown, R.D. (1999). Adding Linguistic Knowledge to a Lexical Example-Based Translation System. In *Proceedings of the Eighth International Conference on Theoretical and Methodological Issues in Machine Translation* - TMI-99, pp. 22-32. Chester, UK, August.

Brown, R.D.; Frederking, R. (1995). Applying Statistical English Language Modeling to Symbolic Machine Translation. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation* - TMI'95, pp. 221-239. Leuven, Belgium, July 5-7.

Carlson, L.; Nirenburg, S. (1992). World Modeling for NLP. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*. Trento, Italy.

Chomsky, N.A. (1957). *Syntatic structures*. Mouton, Hague.

Chomsky, N.A. (1981). *Lectures on Government and Binding*. Dordrecht, Foris.

Dorr, B.J. (1992). The use of Lexical Semantics in Interlingual Translation. In *Journal of Machine Translation*, 7:3, pp. 135-193.

Dorr, B. J. (1993). *Machine Translation: A View from the Lexicon*. The MIT Press, Cambridge.

Dorr, B.J.; Jordan, P.W.; Benoit, J.W. (2000). A Survey of Current Paradigms in Machine Translation. In M. Zelkowitz (ed), *Advances in Computers*, Vol 49, pp. 1-68. Academic Press, London.

Farwell, D.; Wilks, Y. (1991). Ultra: A Mulit-lingual Machine Translator. In *Proceedings of MT Summit III*, Washington.

Fillmore, C. (1968). The case for case. In Bach, E. and Harms, R.T. (orgs.), *Universals in linguistic theory*, pp. 1-88. Rinehard and Winston, New York.

Hutchins, J. (1998). Translation Technology and the Translator. In *Machine Translation Review*. Norfolk.

Isabelle, P. (1987). Machine Translation at the TAUM Group. In *Machine Translation: The State of the Art*, pp. 247-318. Edinburgh University Press, Edinburgh.

Jackendoff, R. (1990). *Semantic Structures*. The MIT Press, Cambridge.

Johnson, R.L.; Whitelock, P. (1987). Machine Translation as an Expert Task. In S. Nirenburg, ed., *Machine translation – Theoretical and methodological issues*, pp. 136-144. Cambridge University Press, Cambridge.

Joshi, A. K. (1987). Introduction to Tree Adjoining Grammar. In A. Manaster Ramer (ed.), *The Mathematics of Language*, pp. 87-114. J. Benjamins.

Kaplan, R.M.; Bresnan, J. (1982). Lexical-Functional Grammar: A Formal System for Grammatical Representation. In Joan Bresnan (ed.), *The Mental Representation of Grammatical Relations.* The MIT Press, Cambridge.

Kay, M. (1994). Machine Translation: The Disappointing Past and Present. In *Survey of the State of the Art in Human Language Technology*. Xerox Palo Alto Research Group, California.

Knight, K. (1999). *A Statistical MT Tutorial Workbook. In Johns Hopkins University 1999 Summer Workshop on Language Engineering*, Center for Speech and Language Processing, Baltimore.

Mateus, M.H.M. (1995). Tradução automática: um pouco de história. In M. H. M. Mateus e A. H. Branco (orgs.), *Engenharia da Linguagem*, pp. 115-120. Edições Colibri, Lisboa.

Mitamura, T.; Nyberg, E.H.; Carbonell, J.G. (1991). An Efficient Interlingua Translation System for Multi-lingual Document Production. In *Proceedings of Machine Translation Summit III*, Washington D.C, July 2-4.

Nirenburg, S. (1987). Knowledge and choices in machine translation. In *Machine translation – Theoretical and methodological issues*, pp. 1-15. Cambridge University Press, Cambridge.

Nirenburg, S.; Raskin, V.; Tucker, A.B. (1987). The structure of interlingua in TRANSLATOR. In *Machine translation – Theoretical and methodological issues*, pp. 90-113. Cambridge University Press, Cambridge.

Nirenburg, S.; Goodman, K. (1991). *The KBMT Project: A case study in Knowledge-Based Machine Translation*. Morgan Kaufmann Publishers, California.

Nyberg, E.H.; Mitamura, T. (1992). The Kant System: Fast, Accurate, High-quality Translation in Practical Domains. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING '92)*, Nantes.

Och, F.J.; Weber, H. (1998). Improving Statistical Natural Language Translation with Categories and Rules. *In COLING '98*.

Rosseta, M.T. (1994). *Compositional Translation*. Kluwer Academic Publishers. Dordrecht, The Netherlands.

Santos, D. (1998). Um olhar computacional sobre a tradução. In *Revista Internacional de Língua Portuguesa*.

Schank, R. (1975). *Conceptual Information Processing*. North-Holland Publishing Company.

Slocum, J. (1985). A Survey of Machine Translation: Its History, Current Status, and Future Prospects. In J. Slocum (org.), *Machine Translation Systems*, pp.1-41. Cambridge University Press, Cambridge.

Uchida, H.; Zhu, M.; Senta, T.D. (1999)[*]. *The UNL, a Gift for a Millennium*. UNU/IAS/UNL Center. Tokyo.

UNL (2001)[*]. *The Universal Networking Language (UNL) Specifications*. UNU/IAS/UNL Center. Tokyo.

Uschold, M.; Gruninger, M. (1996). Ontologies: principles, methods and applications. In *The Knowledge Engineering Review*, Vol. 11:2, pp. 93-136.

Vogel, S.; Och, F.J.; Ney, H. (2000). The Statistical Translation Module in the Vermobil System. In *KOVENS*.

Vogel, S.; Och, F.J.; Tillmann, C.; Nieben, S.; Sawaf, H.; Ney, H. (2000b). Statistical Methods for Machine Translation. *In Verbmobil: Foundations of Speech-to-Speech Translation* pp. 377-393. Wolfgang Wahlster (ed.). Springer Verlag, Berlin.

Wahlster, W. (1993). Verbmobil, translation of face-to-face dialogs. In *Proceedings of the Fourth Machine Translation Summit*, pp. 127-135. Kobe.

Ward, N. (1999). Machine Translation. Chapter 20 of *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* by Jurafsky, D. and Martin, J.H., Prentice-Hall. Wong, S K.

[*] Disponíveis em http://www.unl.ias.unu.edu/