Arquitectura do sistema de tradução da Logos e sua aplicação no desenvolvimento do par inglês-português

---

Anabela Barreiro Colasuonno	Linguista Inglês-Português
Logos Corporation	Tel. (geral) + 1 – 800 – 564 6768
100 Enterprise Drive	Tel. (directo) +1 – 973 – 366 7225, ext. 142
Suite 501	Fax +1 – 973 – 366 7697
Rockaway, NJ 07856	E-mail: anabela@logos-usa.com
USA	http://www.logos-usa.com

Diariamente testemunhamos avanços tecnológicos que permitem encurtar o tempo e a distância entre os habitantes da aldeia global em que vivemos. Os computadores apresentam-se como ferramentas importantes em todas as áreas da comunicação, desempenhando um papel relevante nas designadas "*indústrias da língua*".

No campo da tradução, o computador é cada vez mais utilizado. A construção de ferramentas linguísticas e a utilização de novas técnicas emergem velozmente. A "*globalização*" manifestada através da abertura de novas áreas de negócios a mercados multilíngues impõe o desenvolvimento de meios que permitem o fácil e rápido acesso aos produtos.

A tradução, tal como outras actividades que requerem inteligência, é um processo baseado no conhecimento (Whitelock & Kilby, 1995). Para este processo existem vários produtos de software que auxiliam a tarefa dos tradutores: dicionários, terminologias, glossários electrónicos e memórias de tradução, etc. Existem, para além destes, ferramentas mais complexas, integrando as primeiras, que permitem analisar e traduzir frases e textos, minimizando os esforços humanos. A história deste tipo de ferramentas iniciou-se há cerca de trinta anos[1]. O trabalho tem sido árduo e os resultados estão ainda longe de serem perfeitos. Uma tradução de qualidade, em particular de textos não especializados, permanece até hoje como um objectivo ambicioso. Mas apesar da consciência das dificuldades deste tipo de tarefa, o interesse não se perdeu e actualmente há muitas empresas a investir na área da tradução automática, especialmente de textos técnicos (Barreiro-Colasuonno et al., 1996). Apesar da tradução automática reduzir o tempo e os esforços do tradutor, o papel deste continua a ser fulcral no que diz respeito ao controle da qualidade. A pós-edição é uma tarefa indispensável.

O meu contributo neste seminário deve-se ao facto de participar no desenvolvimento prático de um sistema de tradução automática numa empresa - Logos Corporation - que, desde há 30 anos trabalha nesta área. O objectivo da minha comunicação consiste em apresentar dados acerca do projecto de desenvolvimento de um novo par de línguas, neste caso o de inglês-português, e descrever a arquitectura do sistema de tradução da Logos no qual este desenvolvimento assenta.

DESENVOLVIMENTO DO PAR DE TRADUÇÃO INGLÊS-PORTUGUÊS

O desenvolvimento de um sistema comercial de tradução automática passa pela criação de múltiplos componentes, tais como dicionários e conjuntos de regras, filtros, interfaces e programas que permitem a transferência gradual da informação de um texto na língua-fonte de forma a gerar resultados linguísticos aceitáveis na língua-alvo.

A Logos tem aplicado esta tecnologia aos pares de línguas inglês-alemão, inglês-francês, inglês-italiano e inglês-espanhol, por um lado, e alemão-inglês, alemão-francês e alemão-italiano, por outro. Uma das prioridades actuais consiste no desenvolvimento de novos pares de línguas, nomeadamente inglês-português. Este projecto conta actualmente com o apoio da Fundação para a Ciência e Tecnologia, através da colaboração de quatro bolseiros que, durante um estágio de seis meses, trabalham no nosso Centro de Desenvolvimento Tecnológico. De momento, as principais actividades do projecto são:

- desenvolvimento da componente lexical através da elaboração de um dicionário informatizado inglês/português, tendo em conta as informações gramaticais e semânticas necessárias à tradução;

[1] Uma descrição dos sistemas iniciais pode ser encontrada em Slocum (1985).

- desenvolvimento de uma base de regras semânticas que expressam correspondências subtis entre o significado de expressões em inglês e em português, tais como *raise(vt) N(child,animal,etc)=criar* e *raise(vt) N(acc-misc.vegetative)= cultivar*;
- desenvolvimento de ferramentas de criação semi-automática de glossários e bases de terminologia;
- desenvolvimento do processo de análise e geração.

O nosso objectivo a curto prazo consiste em lançar no mercado um produto que permita auxiliar a tarefa da tradução de inglês para português. Este lançamento está previsto para meados do ano 2000.

SISTEMA DE TRADUÇÃO AUTOMÁTICA DA LOGOS

O sistema de tradução automática da Logos é um sistema híbrido, alcançado através de uma representação intermédia, combinando características dos sistemas baseados em regras de transferência e dos sistemas baseados numa interlíngua (Tucker, 1987 e Hutchins & Somers, 1992). É um sistema modular multilíngue, de orientação semântico-sintáctica, com base num modelo do processo mental do tradutor humano (Scott, 1989). A representação da linguagem natural integra a sintaxe e a semântica em todos os níveis de processamento da frase.

No sistema Logos, os elementos da linguagem natural são representados através de uma linguagem abstracta com classes que representam as propriedades semânticas e sintácticas das palavras. Esta linguagem abstracta é designada de "*semantico-syntactic abstraction language*" (SAL). O SAL é, assim, um sistema de análise e organização sintáctico-semântica do léxico de uma língua (Scott, 1989). Cada vez mais especialistas consideram que "ontologias" ou "metalinguagens" como o SAL desempenham um papel fundamental no processamento automático das línguas naturais.

Com cerca de 1.000 categorias, SAL é uma hierarquia constituída por quatro níveis de abstracção: o nível sintáctico (categoria gramatical) e três níveis de conceito abstracto, designados de superconjunto (superset), conjunto (set) e subconjunto (subset). Essas categorias compreendem tanto classificações gramaticais (verbo bitransitivo, nome próprio, etc.) como classificações conceptuais (método, instrumento, objectivo, etc.) que estão sistematicamente relacionadas entre si (por exemplo, uma sub-classificação em termos semânticos da categoria dos verbos transitivos).

A figura 1 mostra um fragmento da ontologia SAL para a classe dos substantivos em inglês.

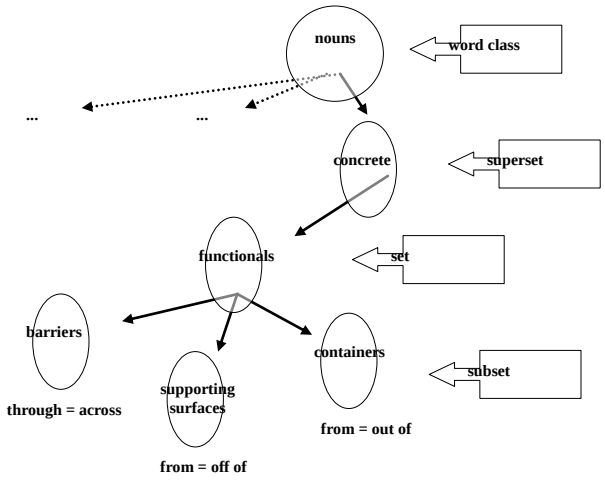


Fig. 1 - SAL - Ontologia para os substantivos em inglês

No início do processo de tradução, numa fase de procura no dicionário (*dictionary lookup*), a frase da língua-fonte é transformada numa série de elementos SAL e, em seguida, é comparada com um conjunto de padrões linguísticos (regras semânticas e sintácticas) com os quais se podem combinar.

O processo de tradução é realizado progressivamente em seis módulos (RES, Tran1, Tran2, Tran3, Tran4 e Generate). A resolução de ambiguidades homógrafas e segmentação da frase em orações é feita no primeiro módulo (RES). A criação de nós apropriados de uma análise gramatical de baixo para cima (*bottom-up parse*) é realizada em quatro níveis de análise (Tran 1-4). Tran1 e 2 correspondem aos nós de nível mais baixo (sintagmas nominais) enquanto que Tran 3 e 4 correspondem aos nós de nível mais alto (sintagmas verbais e frases). A expansão dos nós em cada módulo nas estruturas frásicas apropriadas da língua-alvo opera-se nos quatro níveis de Tran. O último módulo (*Generate*) corresponde à geração da frase da língua-alvo.

A figura 2 representa a implementação do modelo linguístico da Logos.

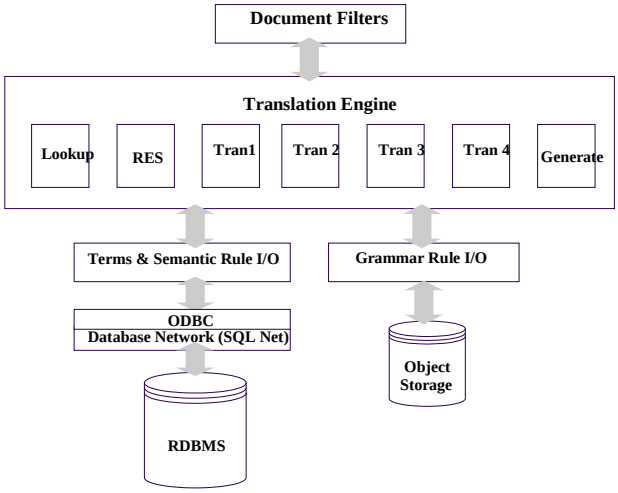


Fig. 2 - Mecanismo de tradução.

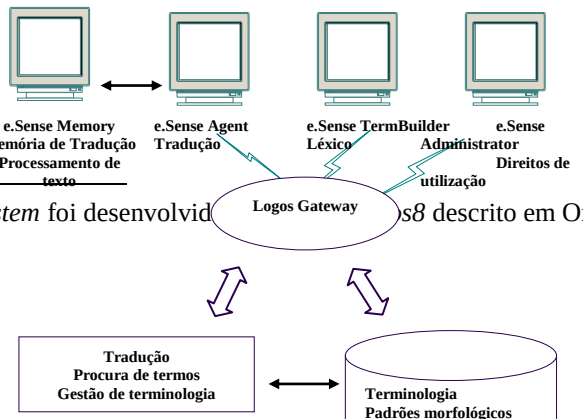
Interfaces

O dicionário está implementado numa base de dados relacional (RDBMS – *Relational Database Management System*) com interfaces para uso cliente-servidor em Intranets ou na Internet e inclui ferramentas sofisticadas de armazenamento, organização e manuseamento de terminologias e dicionários, tais como ferramentas de extracção de termos e importação e exportação de bases terminológicas externas. A comunicação é feita através de uma porta de entrada comum (*gateway*). As três interfaces de cliente fazem parte da nova geração de produtos chamados *e.Sense Enterprise System*[2] e correspondem a:

- *e.Sense Agent* – utilizado para a submissão e recolha de terminologia e trabalhos de tradução;
- *e.Sense TermBuilder* – utilizado para gestão do léxico;
- *e.Sense Administrator* – utilizado para o registo de utilizadores e para direitos de utilização.

O sistema também inclui memórias de tradução (LTM – *Logos Translation Memory*) para reutilizar frases traduzidas e verificadas por um humano e integrar a memória de tradução com a tradução automática. A memória de tradução é baseada no *Translation Manager* da IBM e está integrada com o *e.Sense Agent*.

A figura 3 representa a arquitectura básica de *e.Sense*.



[2] *e.Sense Enterprise System* foi desenvolvid *s8* descrito em Orliac (1998).

Fig. 3 - Arquitectura do sistema Logos – e.Sense

Ferramentas de criação e manutenção semi-automatizada das bases de dados lexical e semântica

Das três interfaces atrás referidas, *e.Sense TermBuilder* desempenha um papel de relevo na fase de desenvolvimento da base de dados lexical, ao permitir a codificação automática de cada entrada (Orliac & Arrieta, 1999). Com a expansão do sistema a novos pares de línguas, tais como o de inglês-português, surge a necessidade de criar novos métodos de reutilização dos dados existentes, resultando daí o desenvolvimento de novas funcionalidades dentro de TermBuilder. Por exemplo, o acréscimo de palavras a partir do texto-fonte, é feito através da pesquisa de termos não encontrados (*Term Search*). Através de regras expressas num formalismo lógico, estas novas entradas lexicais podem ser identificadas e classificadas com atributos gramaticais como: género e número, identificação dos modificadores de um termo com mais do que um elemento, identificação do padrão morfológico e outros (*AutoCode*).

TermBuilder foi criado a pensar nas necessidades dos clientes. Estes podem criar os seus próprios dicionários e modificar um ou mais atributos de uma entrada. As modificações podem ser aplicadas a uma entrada específica, a todas as ocorrências da palavra-fonte e a todas as ocorrências da palavra-alvo.

Outras duas funcionalidades importantes são *Import* e *Export*. *Import* permite aos utilizadores acrescentar entradas (substantivos, adjectivos, advérbios e verbos) a partir de um documento escrito na língua-fonte ou a partir de um glossário bilingue. Listas de terminologia podem ser actualizadas automaticamente (em grupos) ou semi-automaticamente (interactivamente). Para importar automaticamente glossários já existentes para a base de dados da Logos, basta que se apresente o código respeitante ao par de línguas, à empresa, à área temática dos termos, o termo fonte e a sua categoria gramatical, o género do termo fonte, o termo alvo e o género do termo alvo. Todos os outros atributos de cada entrada serão gerados automaticamente. *Export* permite aos utilizadores exportar listas de terminologia (palavras ou orações, entradas longas, palavras com letras acentuadas ou hífens, etc.). O resultado deste tipo de operações pode ser visto em relatórios pré-definidos, adequados às necessidades ou preferências do operador. É possível, para além disso, definir critérios de filtragem, de modo a seleccionar as entradas por código da empresa, código da área temática, utilizador, data, etc.

Uma especificidade da aproximação da Logos é a construção e manutenção sistemática da componente de resolução de ambiguidades lexicais a que chamamos "*Semtab Database*". No que diz respeito ao conhecimento de base de dados semântica, está em construção uma ferramenta (*Semantha*) que permite gerar automaticamente regras que adaptam a selecção de transferências lexicais às necessidades de contextos específicos. Esta ferramenta permite minimizar a intervenção do linguista ou do tradutor. A desambiguação efectua-se com base em propriedades semântico-sintácticas dos elementos que compõem a estrutura. Existem cerca de 11 mil regras já implementadas para os vários pares de línguas. Com a nova ferramenta, o linguista ou tradutor apenas necessita de documentar as correspondências frásicas entre duas línguas numa linha de comentário. Todas as transformações internas ao sistema serão efectuadas automaticamente. A eficácia de *Semantha* conduzirá a um melhor processamento e geração e, consequentemente, a uma melhoria significativa dos resultados.

CONCLUSÃO

Os sistemas de tradução automática já provaram auxiliar grandemente o trabalho do tradutor e, em casos de grandes volumes de tradução, apresentam-se como a única solução viável. A necessidade de continuar a investigação neste campo é enorme, mas ao mesmo tempo que os especialistas na área procuram responder às exigências de qualidade e performance, as aplicações actuais continuam a ter cada vez mais sucesso e, a tradução de milhões de páginas de texto por uma máquina, representa um grande passo em frente.

Com as novas tecnologias que a Logos tem vido a desenvolver é possível traduzir cerca de 130 mil palavras por hora. Com ferramentas que tornam este desenvolvimento mais eficiente, a ênfase agora é dada ao aumento da quantidade e qualidade das bases de dados nas línguas já existentes e a extensabilidade a novos pares de línguas, como é o caso do português. Os instrumentos novos permitem a aquisição de novos léxicos para tradução automática, mas podem ser igualmente extensíveis a outras aplicações.

REFERÊNCIAS

Barreiro-Colasuonno, A., Wittmann, L. & Pereira, M. J. (1996). "Lexical differences between European and Brazilian Portuguese". In *The INESC Journal of Research & Development*, vol. 5, n°2, Lisboa, Jan-Dez.

Hutchins, W. J. & Somers, H. L. (1992). *An Introduction to Machine Translation*. Academic Press, London.

Orliac, B. (1998). "The Logos8 System". In *Machine Translation and the Information Soup*. Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas, AMTA'98, Springer.

Orliac, B. & Arrieta, K. (1999). "TermBuilder: A Lexical Knowledge Acquisition Tool for the Logos Machine Translation System". In Proceedings of the 7th Machine Translation Summit VII'99, 593-598, Sept., Singapore.

Scott, B. E. (1989). "The Logos System". In Proceedings of MT Summit II, Munich.

Slocum, J. (1985). "A Survey of Machine Translation: its History, Current Status, and Future Prospects". In *Computational Linguistics* 11:1-17.

Tucker, A. B. (1987). "Current Strategies in Machine Translation Research and Development". In *Machine Translation: Theoretical and Methodological Issues*. Ed. Sergei Nirenburg. Cambridge University Press.

Whitelock, P. & Kilby, K. (1995). *Linguistic and Computational Techniques in Machine Translation System Design*. 2ª edição, UCL Press, London.