Tradução Automática Baseada em Exemplos

Alberto Manuel Brandão Simões
ambs@di.uminho.pt

Simpósio Doutoral, Departamento de Informática, 2006

1 Identificação

- **Orientadores:**
  José João Antunes Guimarães Dias de Almeida
  Departamento de Informática, Universidade do Minho
  jj@di.uminho.pt

  Diana Maria de Sousa Marques Pinto dos Santos
  Linguateca, Sintef, Oslo
  diana.santos@sintef.no

- **Data de Início:**
  21 de Setembro de 2004

- **Data Término:**
  21 de Setembro de 2007

2 Resumo

- **Área de Investigação e Desenvolvimento:**
  Processamento de Linguagem Natural

- **Resumo:**
  A tradução automática baseada em exemplos (EBMT) é suportada essencialmente pela grande quantidade de frases ou segmentos já traduzidos (exemplos). Durante o processo de tradução, o sistema tenta encontrar nesses exemplos porções do texto a traduzir, e construir a tradução.

  Com este trabalho pretende-se estudar de que forma a língua portuguesa se adapta a este tipo de tradução, e que ferramentas podemos usar para produzir exemplos mais adequados a esta língua. Será ainda criado um protótipo de tradutor, bem como

propostas de metodologias de avaliação de dicionários probabiliticos de tradução, e de tradução automática.

3 Contribuições

- **Publicações:**
  - Alberto Simões e José João Almeida, "Combinatory Examples Extraction for Machine Translation", no prelo.

    One of the bottlenecks of example-based machine translation (EBMT) is to be able to amass automatically quantities of good examples.

    In our work in EBMT, we are investigating how far one can go by performing example extraction from parallel corpora using Probabilistic Translation Dictionaries to obtain example segmentation points.

    In fact, the success of EBMT highly depends on examples quality and quantity, but also in their length. Thus, we give special importance on methods to extract different size examples from the same translation unit.

  - Alberto Simões e José João Almeida, "NATools — A Statistical Word Aligner Workbench", Socidedade Española para el Procesamiento del Lenguaje Natural, Sep 2003.

    This document presents the TerminUM project and the work done in its statistical word aligner workbench (NATools). It shows a variety of alignment methods for parallel corpora and discusses the resulting terminological dictionaries and their use: evaluation of sentence translations; construction of a multi-level navigation system for linguistic studies or statistical translations.

  - Alberto Simões, Xavier Gómez Guinovart e José João Almeida, "Distributed Translation Memories implementation using WebServices", Sociedade Española para el Procesamiento del Lenguaje Natural, Jul 2004.

    Translation Memories are very useful for translators but are difficult to share and reuse in a community of translators. This article presents the concept of Distributed Translation Memories, where all users can contribute and sharing translations. Implementation details using WebServices are shown, as well as an example of a distributed system between Portugal and Spain.

4 Enquadramento

- **Enquadramento Científico:**
  A tradução automática é, sem dúvida, uma das áreas principais de investigação no processamento computacional da linguagem natural. Durante os cerca de 55 anos de investigação nesta área, que se tem abordado este problema de formas diferentes. Inicialmente pretendia-se a criação de um sistema de tradução entre um par de línguas, o que levou ao desenvolvimento de sistemas monolíticos cujo código escondia todo um conjunto de regras e de léxico usado na tradução. Seguiu-se a separação do código da informação linguística. No entanto as regras necessárias para a tradução eram demasiado complexas, e dependentes da língua.

  Na tentativa de solucionar este problema surgiram abordagens baseadas em corpora paralelos: grandes quantidades de texto e respectivas traduções. Estas abordagens tentam, de forma automática, extrair informação de como se realiza a tradução entre estas duas línguas.

  A Tradução Automática Baseada em Exemplos é uma destas abordagens.

- **Motivação:**
  Embora exista investigação na área da tradução automática, pouca ou nenhuma é aquela que inclui a língua portuguesa. De facto, os poucos projectos de tradução automática que estudam a língua portuguesa vêm-na como uma língua de destino, normalmente na sua variante brasileira (p.ex. projecto Apertium).

  No entanto, mesmo os projectos que de alguma forma têm em conta a nossa língua, usam abordagens baseadas em regras, pelo que nos parece imprescindível um estudo da adaptabilidade da língua portuguesa a esta metodologia de tradução.

  Embora se pretenda que a EBMT seja completamente independente de língua cada vez mais se chega à conclusão que tal não é possível.

- **Trabalhos Alternativos:**

  – GIZA++: Franz Josef Och, Hermann Ney. "A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics, volume 29, number 1, pp. 19–51 March 2003

    Ferramenta pública para a aprendizagem de modelos estatísticos para tradução automática. Enquadra-se no uso de corpora de forma estatística para tradução automática.

  – "The Pangloss Mark III Machine Translation System", A Joint Technical Report by NMSU CRL, USC ISI and CMU CMT. Sergei Nirenburg, editor, April 1995. Issued as CMU tech report CMU-CMT-95-145

    O Pangloss é uma ferramenta de tradução automática baseada em exemplos que tem vindo a ser desenvolvida na Universidade do Estado do Novo México,

Universidade da Califórnia do Sul, e Universidade Carnegie Mellon. Embora a equipa tenha vindo a publicar artigos científicos sobre a ferramenta esta não se encontra disponível (aparentemente, nem comercialmente).

- **Bibliografia Principal:**
  - Michael Carl e Andy Way (Eds), "Recent Advances in Example-Based Machine Translation", Serie on Text, Speech and Language Technology, Kluwer Academic Publishers, 2003.
  - I. Dan Melamed, "Empirical Methods for Exploiting Parallel Texts", MIT Press, 2001.

5 Desenvolvimento

- **Macro-planeamento das Actividades:**
  Trabalho realizado:
  - Setembro 2004 a Março de 2005: estado da arte e contextualização na área;
  - Janeiro 2004 a Março de 2005: experiências no melhoramento dos dicionários probabilísticos de tradução por meio de lematização de palavras, e uso das respectivas categorias gramaticais;
  - Fevereiro 2004: criador de classes de palavras de acordo com o contexto em que são usados.
  - Março de 2005 a Julho de 2005: desenvolvimento de um servidor eficiente de unidades de tradução e de dicionários probabilísticos de tradução;
  - Maio de 2005 a Julho de 2005: desenvolvimento de clientes web para o servidor descrito anteriormente;
  - Agosto de 2005 a Dezembro de 2005: desenvolvimento de um extractor de exemplos baseado em corpora paralelos, por meio de um alinhamento ao bloco;
  - Dezembro de 2005 a Janeiro de 2006: experiências de armazenamento eficiente de exemplos em bases de dados relacionais;
  - Janeiro de 2006: escrita de vários artigos para workshops e conferências: LREC, EACL, EAMT, Propor.

  Plano futuro:
  - alinhamento do COMPARA ao bloco, e sua disponibilização na WEB;
  - desenvolvimento de um protótipo de tradutor baseado em exemplos;
  - desenvolvimento de um sistema de tradução de segmentos *on-the-fly*;
  - proposta de avaliação de dicionários probabilísticos de tradução;

- inferidor de regras de tradução;
- análise da possibilidade do uso de paráfrases para a tradução automática e possível uso na avaliação de tradução automática;
- proposta de avaliação de tradutores automáticos;
- escrita da dissertação.

- **Recursos Necessários:**
  Além da leitura de outra bibliografia não referida anteriormente, e que aqui também não será detalhada a investigação na área da tradução automática baseada em exemplos requer:
  - grandes quantidades de corpora paralelos alinhados à frase de onde os exemplos de tradução, dicionários probabilísticos e terminologia possa ser extraída automaticamente;
  - grande poder computacional, nomeadamente em termos de memória do sistema, para o armazenamento eficiente dos exemplos durante o seu uso, e em termos de velocidade de CPU, já que a extracção de dicionários probabilísticos, e a extracção de exemplos são processos exigentes. Nesse sentido, pretende-se estudar a possibilidade do uso do Search para ajudar nesta tarefa.

- **Recursos Disponibilizados:**
  Actualmente já existem disponíveis os seguintes recursos:
  - NATools: pacote de ferramentas para manuseamento de corpora paralelos que inclui, entre outras ferramentas:
    * alinhador à frase;
    * extractor de dicionários probabilisticos de tradução;
    * extractor de exemplos;
    * servidor de corpora e dicionários de tradução;
    * conjunto de CGIs para a consulta dos recursos produzidos;
  - Corpora paralelos disponíveis via web em `http://linguateca.di.uminho.pt/albin/nat` e em `http://eremita.di.uminho.pt/albin/nat`:
    * Parlamento Europeu (PT–EN) ≈ de 1,000,000 UT[1], 30M palavras;
    * Parlamento Europeu (PT–ES) ≈ de 1,000,000 UT, 30M palavras;
    * Parlamento Europeu (PT–FR) ≈ de 1,000,000 UT, 30M palavras;
    * Constituição Portuguesa (PT–EN) ≈ de 2,700 UT;
    * Constituição Portuguesa (PT–FR) ≈ de 2,700 UT;
    * Ficheiros de I18n (PT–EN) ≈ de 630,000 UT, 3M palavras;
    * ClavisIII (PT–LA) ≈ de 2,700 UT;

[1] UT — Unidades de Tradução

Pretende-se disponibilizar até ao fim do doutoramento:

- outros corpora que venham a surgir;
- o pacote NATools mais robusto e com mais ferramentas, incluindo um protótipo de tradutor baseado em exemplos;

6 Avaliação

- **Análise Comparativa:**
  O trabalho que tem vindo a ser desenvolvido baseia-se em trabalho já realizado mas para outras línguas (uma vez que estamos interessados no Português), com algumas diferenças nos algoritmos propostos, bem como na abordagem realizada.

- **Auto-avaliação do Trabalho Realizado:**
  O trabalho tem vindo a evoluir de forma algo lenta para o planeado inicialmente, mas ser por em causa os objectivos que se pretende atingir.

- **Auto-avaliação da Documentação Produzida:**
  A documentação produzida até ao presente momento divide-se em três grandes blocos:

  - artigos submetidos sobre vários pontos do trabalho realizado;
  - documentação do código e do uso das ferramentas desenvolvidas;
  - enquadramento do trabalho na história da tradução automática, e contextualização das técnicas habitualmente usadas;

7 Dificuldades

- **Dificuldades Técnico-Científicas:**
  O manuseamento de grandes quantidades de texto é complicado, nomeadamente quando a maior parte das ferramentas (como bases de dados) se contentam com a indexação à chave, sem procura textual eficiente. Nesse sentido, tem-se vindo a desenvolver software para resolver de forma eficiente estes problemas.