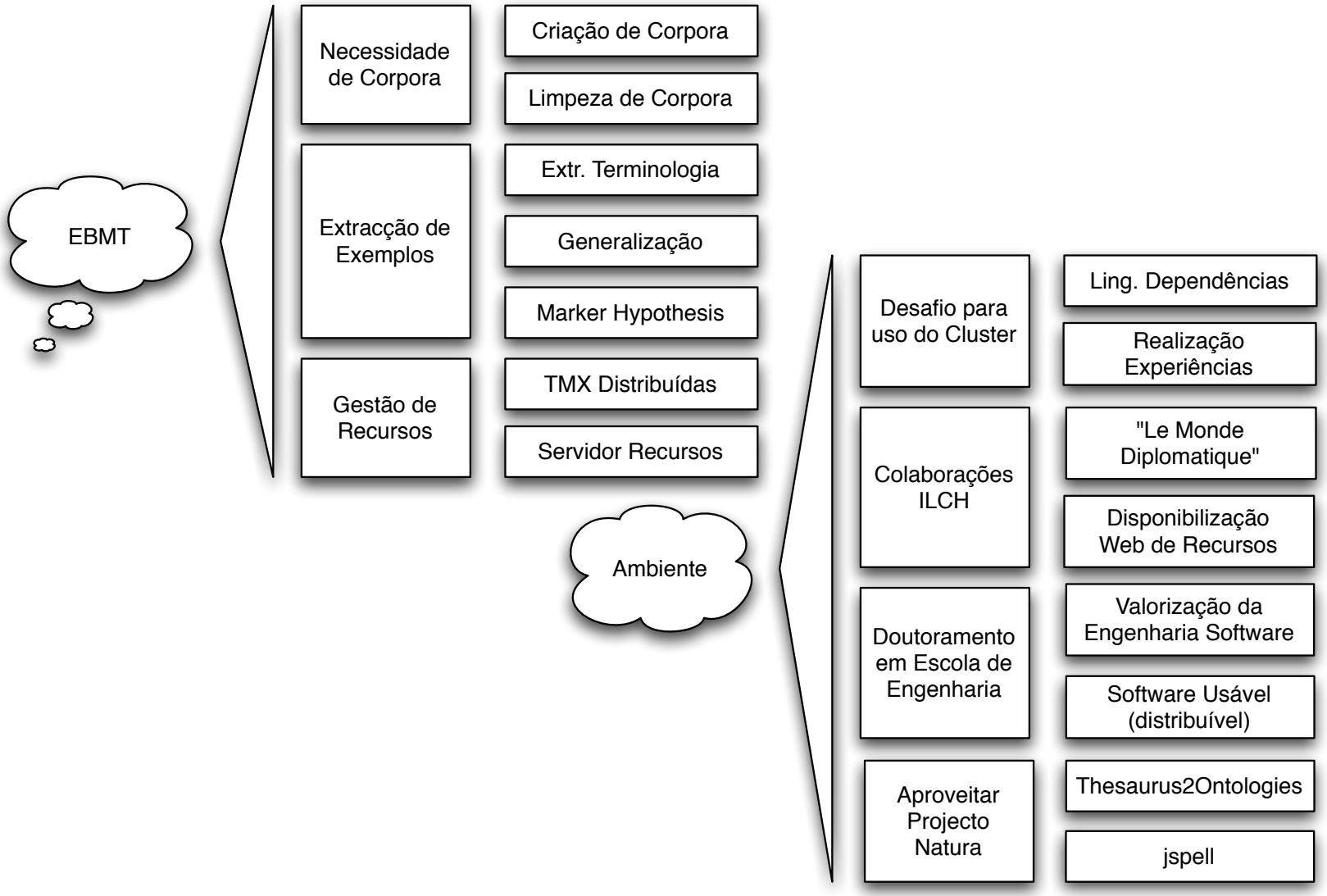
Extracção de Recursos de Tradução com base em Dicionários Probabilísticos

Alberto Manuel Brandão Simões
ambs@di.uminho.pt

Orientação
José João Almeida

Simpósio Doutoral da Linguateca 2007a

Visão Cronológica



Extracção de Recursos Bilingues de Tradução

A Tradução (Automática ou Assistida) usa:

- Dicionários de Tradução $\subset$ *PTDs*;
- Unidades de Tradução $\equiv$ *Corpora*;
- Exemplos de Tradução $\equiv \mathcal{P}(Corpora)$;
- Terminologia Bilingue;
- Regras de Tradução;
- Classes de Palavras;

1. Pipeline de processos:
   1. Criação e processamento de Corpora;
   2. Alinhamento de Corpora;
   3. Extracção de Relacionamentos Bilingues (PTDs);
   4. Extracção de Exemplos de Tradução;
   5. Extracção de Terminologia Bilingue;
   6. Generalização de Exemplos;

Extracção de Recursos Bilingues de Tradução

1. Pipeline de processos:
   1. Criação e processamento de Corpora;
   2. Alinhamento de Corpora;
   3. Extracção de Relacionamentos Bilingues (PTDs);
   4. Extracção de Exemplos de Tradução;
   5. Extracção de Terminologia Bilingue;
   6. Generalização de Exemplos;
2. Foi preciso ter em consideração:
   1. Disponibilização das ferramentas;
   2. Disponibilização/Validação dos recursos;
   3. Manuseamento eficiente de Recursos;
   4. Processamento de grandes quantidades;

- Métodos estatísticos precisam de grandes quantidades de corpora.
- Qualidade dos resultados depende muito da qualidade dos corpora.

- Métodos estatísticos precisam de grandes quantidades de corpora.
- Qualidade dos resultados depende muito da qualidade dos corpora.

- Há corpora... mas:
  - EuroParl: Grande mas com ruído (11 línguas);
  - Acquis-JRC: Médio, menos ruidoso (20 línguas);
  - COMPARA: Pequeno, literário (PT:EN);
- Foi realizado trabalho em:
  - Criação de Corpora:
    - Le Monde Diplomatique: pequeno, ruído mínimo (FR:PT);
    - EurLex: monstruoso, ruidoso, rico em terminologia (Milhões de UTs, PT:EN/PT:FR/PT:ES);
  - Tratamento e filtragem de Corpora: TMX::TMX [AS07];

- Métodos estatísticos precisam de grandes quantidades de corpora.
- Qualidade dos resultados depende muito da qualidade dos corpora.

- Há corpora... mas:
  - EuroParl: Grande mas com ruído (11 línguas);
  - Acquis-JRC: Médio, menos ruidoso (20 línguas);
  - COMPARA: Pequeno, literário (PT:EN);
- Foi realizado trabalho em:
  - Criação de Corpora:
    - Le Monde Diplomatique: pequeno, ruído mínimo (FR:PT);
    - EurLex: monstruoso, ruidoso, rico em terminologia (Milhões de UTs, PT:EN/PT:FR/PT:ES);
  - Tratamento e filtragem de Corpora: TMX::TMX [AS07];

- Obtiveram-se corpora de diferentes géneros e diferentes línguas;
- Criou-se um mecanismo automático eficaz de limpeza de corpora;

- Nem todos os corpora estão alinhados;
- Sem alinhamento à frase os corpora não são produtivos;

- Nem todos os corpora estão alinhados;
- Sem alinhamento à frase os corpora não são produtivos;

- alinhamento à frase;
- conjunto de scripts de alinhamento:
  - baseado no easy-align (IMS Workbench);
  - injectores de HTML, PDF, etc.
- alguma reflexão sobre avaliação de alinhadores [SA07a];

- Nem todos os corpora estão alinhados;
- Sem alinhamento à frase os corpora não são produtivos;

- alinhamento à frase;
- conjunto de scripts de alinhamento:
  - baseado no easy-align (IMS Workbench);
  - injectores de HTML, PDF, etc.
- alguma reflexão sobre avaliação de alinhadores [SA07a];

- Desenvolveu-se um mecanismo automático para o alinhamento de HTMLs e PDFs;

- A tradução precisa de dicionários bilingues;
- A criação (manual) de dicionários é morosa;

- A tradução precisa de dicionários bilingues;
- A criação (manual) de dicionários é morosa;

- nat-ptd (original) [Sim04]:
  - escalável;
  - eficiente:
    - várias ordens de grandeza mais rápido que GIZA++...
    - ...mas não fazem a mesma coisa!
- experiências de melhoramento dos dicionários:
  - lematização de palavras, ou só verbos;
  - tratamento de unidades multi-palavra;
- trabalho em progresso na avaliação/comparação de dicionários;

- A tradução precisa de dicionários bilingues;
- A criação (manual) de dicionários é morosa;

- nat-ptd (original) [Sim04]:
  - escalável;
  - eficiente:
    - várias ordens de grandeza mais rápido que GIZA++...
    - ...mas não fazem a mesma coisa!
- experiências de melhoramento dos dicionários:
  - lematização de palavras, ou só verbos;
  - tratamento de unidades multi-palavra;
- trabalho em progresso na avaliação/comparação de dicionários;

- Um PTD não é um dicionário;
- Os PTDs são bons para estabelecer âncoras entre unidades de tradução;
- É possível construir um dicionário a partir de um PTD;
- Foi implementada uma álgebra sobre PTDs;

- As unidades de tradução do alinhamento à frase são grandes.
- Segmentos grandes são menos (ou nada) reutilizáveis.

- As unidades de tradução do alinhamento à frase são grandes.
- Segmentos grandes são menos (ou nada) reutilizáveis.

- uso de âncoras e matriz de alinhamento [SA06a]
  - tirar partido da pouca movimentação de palavras na tradução técnica/simultânea;
  - tirar partido das âncoras obtidas nos PTDs;
- uso de Marker Hypothesis (under work):
  - método usado em ferramentas de EBMT (Gaijin/MaTrEx – Andy Way)
  - uso de um chunker superficial;
  - alinhamento de chunks baseado em PTDs;

- As unidades de tradução do alinhamento à frase são grandes.
- Segmentos grandes são menos (ou nada) reutilizáveis.

- uso de âncoras e matriz de alinhamento [SA06a]
  - tirar partido da pouca movimentação de palavras na tradução técnica/simultânea;
  - tirar partido das âncoras obtidas nos PTDs;
- uso de Marker Hypothesis (under work):
  - método usado em ferramentas de EBMT (Gaijin/MaTrEx – Andy Way)
  - uso de um chunker superficial;
  - alinhamento de chunks baseado em PTDs;

- É possível obter exemplos pequenos (sintácticos ou não) usando PTDs;
- A existência de PTDs simplifica o alinhamento de chunks;

- Existem movimentos de palavras que são derivados de regras sintácticas;
- Estas regras sintácticas podem ser expressas formalmente;

- Existem movimentos de palavras que são derivados de regras sintácticas;
- Estas regras sintácticas podem ser expressas formalmente;

- o uso de padrões permite expressar reordenação de palavras[SA07b]
  - desde reordenações simples de substantivo e adjectivo;
  - até frases preposicionais/adjectivais mais complicadas;
- estes padrões representam frases nominais simples;
- a aplicabilidade dos padrões pode ser restringida:
  - apenas com determinadas condições morfológicas;
  - de acordo com predicados genéricos (Perl-based);

- Existem movimentos de palavras que são derivados de regras sintácticas;
- Estas regras sintácticas podem ser expressas formalmente;

- o uso de padrões permite expressar reordenação de palavras[SA07b]
  - desde reordenações simples de substantivo e adjectivo;
  - até frases preposicionais/adjectivais mais complicadas;
- estes padrões representam frases nominais simples;
- a aplicabilidade dos padrões pode ser restringida:
  - apenas com determinadas condições morfológicas;
  - de acordo com predicados genéricos (Perl-based);

- Os padrões são directamente relacionados com regras de tradução;
- Os padrões são dependentes do par de língua;
- A escrita de padrões é guiada pelas regras sintácticas da língua;
- As entradas extraídas são (na sua maioria) terminológicas;

- Embora segmentos pequenos sejam mais reutilizáveis, continuam a ser aplicáveis apenas a determinadas situações;
- A generalização aumenta a aplicabilidade dos exemplos ou entradas terminológicas;

- Embora segmentos pequenos sejam mais reutilizáveis, continuam a ser aplicáveis apenas a determinadas situações;
- A generalização aumenta a aplicabilidade dos exemplos ou entradas terminológicas;

Para a generalização é habitual:

- construção de classes de EMs (especialmente EMs numéricas);
- construção de classes de palavras:
  - usando o contexto (uso de n-grams – Brown et al)
  - por fixação de palavras em entradas terminológicas ([SA07b]);

- Embora segmentos pequenos sejam mais reutilizáveis, continuam a ser aplicáveis apenas a determinadas situações;
- A generalização aumenta a aplicabilidade dos exemplos ou entradas terminológicas;

Para a generalização é habitual:

- construção de classes de EMs (especialmente EMs numéricas);
- construção de classes de palavras:
  - usando o contexto (uso de n-grams – Brown et al)
  - por fixação de palavras em entradas terminológicas ([SA07b]);

- A generalização pode ser feita de forma automática;
- A generalização aumenta a aplicabilidade dos exemplos;
- É possível enriquecer classes usando cálculo de multi-sets;

- Necessidade de colocar várias máquinas com o NATools;
- A compilação do NATools não é simples;
- Foi necessário disponibilizar o código a terceiros;

- Necessidade de colocar várias máquinas com o NATools;
- A compilação do NATools não é simples;
- Foi necessário disponibilizar o código a terceiros;

- usar a abordagem típica GNU para configuração:
  - autoconf/automake/libtool;
  - detecção de bibliotecas;
  - detecção de módulos Perl;
- documentação:
  - da API C e Perl;
  - das aplicações/comandos;
- testar a distribuição:
  - unit-testing do software;
  - unit-testing da documentação;

- Necessidade de colocar várias máquinas com o NATools;
- A compilação do NATools não é simples;
- Foi necessário disponibilizar o código a terceiros;

- usar a abordagem típica GNU para configuração:
  - autoconf/automake/libtool;
  - detecção de bibliotecas;
  - detecção de módulos Perl;
- documentação:
  - da API C e Perl;
  - das aplicações/comandos;
- testar a distribuição:
  - unit-testing do software;
  - unit-testing da documentação;

- Permitiu o uso do NATools por diferentes grupos de investigadores;
- Permitiu facilmente instalar o NATools em várias máquinas DI/Linguateca;

2.2. Disponibilização/Validação dos recursos

- A existência de recursos não disponíveis corresponde a recursos inexistentes;
- Recursos que podem parecer demasiado simples são, muitas vezes, os mais úteis para terceiros;
- A utilização de recursos por terceiros ajuda a validar o trabalho e a aumentar a motivação;

- A existência de recursos não disponíveis corresponde a recursos inexistentes;
- Recursos que podem parecer demasiado simples são, muitas vezes, os mais úteis para terceiros;
- A utilização de recursos por terceiros ajuda a validar o trabalho e a aumentar a motivação;

- Disponibilizar serviços web para consulta de recursos:
  - Pesquisa de Concordâncias;
  - Consulta de dicionários probabilísticos de tradução;
  - Consulta de n-gramas;
- Um serviço integrado.

- A existência de recursos não disponíveis corresponde a recursos inexistentes;
- Recursos que podem parecer demasiado simples são, muitas vezes, os mais úteis para terceiros;
- A utilização de recursos por terceiros ajuda a validar o trabalho e a aumentar a motivação;

- Disponibilizar serviços web para consulta de recursos:
  - Pesquisa de Concordâncias;
  - Consulta de dicionários probabilísticos de tradução;
  - Consulta de n-gramas;
- Um serviço integrado.

- Permitiu a utilização de recursos em aulas a cursos de letras;
- Possibilitou a alguns investigadores do ILCH o desenvolvimento das suas teses;
- Permitiu validar vários erros de processamento que não teriam sido detectados sem a intervenção de terceiros.

- A criação de serviços Web requer respostas rápidas (timeouts);
- A extracção de exemplos e terminologia são processos intensivos de consulta a dicionários e corpora;

- A criação de serviços Web requer respostas rápidas (timeouts);
- A extracção de exemplos e terminologia são processos intensivos de consulta a dicionários e corpora;

- Criação do Nat-Server [SA06b];
- Servidor de PTDs, Concordâncias, n-gramas e meta-informação;
- Suporte para vários corpora e diferentes pares de língua;
- Possibilita uso como biblioteca ou servidor sockets;
- API disponível em C e Perl;

- A criação de serviços Web requer respostas rápidas (timeouts);
- A extracção de exemplos e terminologia são processos intensivos de consulta a dicionários e corpora;

- Criação do Nat-Server [SA06b];
- Servidor de PTDs, Concordâncias, n-gramas e meta-informação;
- Suporte para vários corpora e diferentes pares de língua;
- Possibilita uso como biblioteca ou servidor sockets;
- API disponível em C e Perl;

- Resultou numa ferramenta rápida para ambientes interactivos e batch;
- A ferramenta suporta toda a extracção de exemplos;

- A extracção de exemplos de corpora grandes (2M UTs) é demorada (1 semana);
- O uso de um posto de trabalho para extrair exemplos torna-o não usável;
- O uso desta abordagem para a realização de testes é impensável;

- A extracção de exemplos de corpora grandes (2M UTs) é demorada (1 semana);
- O uso de um posto de trabalho para extrair exemplos torna-o não usável;
- O uso desta abordagem para a realização de testes é impensável;

- O DI tem um Cluster (100 nós Xeon);
- A sua utilização foi motivada por todos os grupos;
- Instalar o NATools foi a parte mais simples;
- Para tirar partido do paralelismo:
  - análise do pipeline de extracção de exemplos;
  - especificação da topologia em linguagem formal;
  - construção de um parser e interpretador da linguagem;
  - criação um escalonador de processos que tire partido do paralelismo [SFA07];

2.4. Processamento de grandes quantidades

- A extracção de exemplos de corpora grandes (2M UTs) é demorada (1 semana);
- O uso de um posto de trabalho para extrair exemplos torna-o não usável;
- O uso desta abordagem para a realização de testes é impensável;

- O DI tem um Cluster (100 nós Xeon);
- A sua utilização foi motivada por todos os grupos;
- Instalar o NATools foi a parte mais simples;
- Para tirar partido do paralelismo:
  - análise do pipeline de extracção de exemplos;
  - especificação da topologia em linguagem formal;
  - construção de um parser e interpretador da linguagem;
  - criação um escalonador de processos que tire partido do paralelismo [SFA07];

- O processo de extracção passou a ser muito mais rápido (6 horas);
- Foi desenvolvida uma DSL independente e reutilizável noutros contextos;

Extremely useful work for the scientific community and for translators.

ACL2007 Reviewer

Extremely useful work for the scientific community and for translators.

ACL2007 Reviewer

The underlying work is highly valuable, and can be useful to many people who want to work with parallel corpora.

ACL2007 Reviewer

Extremely useful work for the scientific community and for translators.

ACL2007 Reviewer

The underlying work is highly valuable, and can be useful to many people who want to work with parallel corpora.

ACL2007 Reviewer

We are very sorry to inform you that the following submission was not selected by the program committee to appear at ACL 2007.

ACL2007 Committee

Algumas Contribuições...

- Pacote integrado de processamento de corpora paralelos;
- Plataforma Cliente-Servidor para disponibilização de recursos;
- Uma API ágil para a programação de experiências sobre CP;
- Extractor de dicionários e de exemplos escalável;
- Corpora variado;
  (filtrado, alinhado, indexado, pesquisável, com PTDs e n-gramas)
- Metodologia para extracção de exemplos;
- Linguagem para a extracção de terminologia bilingue;
- Resmas de exemplos e entradas terminológicas;
  (a precisar de um mecanismo eficiente para os armazenar)
- Metodologia para generalização de entradas terminológicas;
- Linguagem para escalonamento de processos paralelos;

Algum Trabalho Futuro...

- Escrever a dissertação;

  ...se me deixarem...
- Teste dos predicados genéricos sobre terminologia;

  ...implementado há quase um mês, há espera de tempo para ser testado...
- Extracção de exemplos usando a Marker Hypothesis;

  ...e colaborar com o Andy Way com uma versão PT...
- Terminar várias experiências incompletas;

  ...muitas delas precisam apenas de análise de resultados...
- Implementar um serviço para disponibilização de exemplos;

  ...o problema actual prende-se com as quantidades...
- Implementar álgebra de multi-sets sobre classes de palavras;

  ...e incluir informação sobre o T2O para ajudar na conciliação de classes...
- E várias outras tarefas que o JJ tem em mente...

  ...e outras tantas que vão surgir à medida que outras forem concluídas...

Algumas publicações

José João Almeida and Alberto Simões.

XML::TMX — processamento de memórias de tradução de grandes dimensões.
In José Carlos Ramalho, João Correia Lopes, and Luís Carríço, editors, *XATA 2007 — 5ª Conferência Nacional em XML, Aplicações e Tecnologias Aplicadas*, pages 83–93, February 2007.

Alberto Simões and J. João Almeida.

Combinatory examples extraction for machine translation.
In Jan Tore Lønning and Stephan Oepen, editors, *11th Annual Conference of the European Association for Machine Translation*, pages 27–32, Oslo, Norway, 19–20, June 2006.

Alberto Simões and J. João Almeida.

NatServer: a client-server architecture for building parallel corpora applications.
Procesamiento del Lenguaje Natural, 37:91–97, September 2006.

Alberto Simões and José João Almeida.

Avaliação de alinhadores.
In Diana Santos, editor, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press, 2007.

Alberto Simões and José João Almeida.

Using alignment patterns for bilingual terminology extraction.
In *The 31st Annual Conference of the German Classification Society on Data Analysis, Machine Learning, and Applications*, 2007.
forthcoming.

Alberto Simões, Rúben Fonseca, and José João Almeida.

Makefile::parallel dependency specification language.
Unpublished, 2007.

Alberto Manuel Brandão Simões.

Parallel corpora word alignment and applications.
Master's thesis, Escola de Engenharia - Universidade do Minho, 2004.