XMLaligner: Exploração de Corpora Paralelos

Ivo Anjo and David Martins de Matos

L^2F – Laboratório de Sistemas de Língua Falada
INESC-ID Lisboa/Instituto Superior Técnico/Universidade Técnica de Lisboa
Rua Alves Redol 9, 1000-029 Lisboa, Portugal
{ivo.anjo,david.matos}@l2f.inesc-id.pt
http://www.l2f.inesc-id.pt/

Resumo Apresenta-se a aplicação XMLaligner, que permite a consulta e manipulação de corpora multilingues estruturados, sendo o seu principal objectivo o de fornecer uma forma simples, mas poderosa, de navegação num corpus paralelo descrito a múltiplos níveis. Tendo em consideração que um corpus pode ser constituído por milhares de documentos, a aplicação, concebida como um "browser" gráfico, permite, além da consulta simultânea e alinhada de documentos do corpus, para cada par de línguas, efectuar pesquisas no contexto desses ficheiros. A ferramenta potencia a utilização de corpora paralelos em aplicações como tradução automática e projecção de características entre vários conjuntos de línguas.

1 Introdução

No processamento das línguas humanas é necessário o recurso a colecções de textos – corpora – que providenciam material observável, a partir do qual se derivam regras e modelos estatísticos que permitem construir aplicações como processadores sintácticos ou sistemas de tradução automática (que podem ser treinados a partir de corpora paralelos). Os corpora multilingues podem ainda ser úteis em aplicações de lexicografia e na projecção de características, descritas para uma dada língua, para outra onde elas não estão ainda descritas, o que os torna úteis na aquisição de informação linguística sobre a língua alvo.

Uma colecção de corpora paralelos é composta pelos documentos nas várias línguas e pela descrição dos alinhamentos entre as línguas. Apesar de ser possível a codificação destas colecções numa grande variedade de formatos, desde bases de dados até ficheiros textuais, são as representações hierárquicas as que melhor combinam a flexibilidade com a simplicidade. É neste ponto que normas como o SGML [1] e, mais recentemente, o XML [2] assumem um papel importante na codificação de documentos.

A utilização de corpora para os fins expostos pressupõe, contudo, a capacidade de extrair informação relevante, seja automática ou manualmente. Ainda que na análise automática existam processos eficazes para sintetizar vários tipos de informação, na análise manual assumem importância não trivial as limitações humanas, seja na manipulação das grandes quantidades de informação não agregada, que é a natureza habitual de um corpus, seja na gestão dos múltiplos componentes (ficheiros).

1.1 Objectivos

Com vista a facilitar a interacção com colecções paralelas, definiu-se o objectivo de construir uma ferramenta que permitisse não só a gestão de ficheiros das colecções alinhadas, mas também a navegação paralela e a execução de consultas no contexto dos textos alinhados. Um outro objectivo foi o de conseguir fornecer a funcionalidade através de uma interface simples de utilizar.

1.2 Estrutura do documento

Considerando a motivação e objectivos expostos, foi criada a aplicação XMLaligner. A secção 2 descreve o contexto de trabalho previsto para a aplicação, em particular, no que respeita ao processamento computacional da língua baseado em corpora, focando-se os corpora estruturados baseados em XML (secção 2.1) e alguns casos concretos (secção 2.2). A secção 3 apresenta a arquitectura e o desenvolvimento do protótipo da aplicação. Finalmente, na secção 4 apresentam-se algumas conclusões e direcções para evolução.

2 Contexto

A motivação para o trabalho advém das necessidades sentidas na aplicação ao português de algoritmos e dados de análise de discurso da língua inglesa [3]. A falta de recursos linguísticos equivalentes torna impossível a tarefa, pelo que surgiu a ideia de projectar a informação disponível para o inglês através de um corpus paralelo. Este corpus possibilita a extracção de informação linguística para processamento aos níveis da semântica e do discurso, permitindo a aplicação final do algoritmo ao português.

2.1 Corpora XML alinhados

Considerando que a informação estrutural dos textos é importante para a análise de discurso, foi definido como requisito que o corpus de trabalho fosse estruturado e que, se possível, admitisse marcações adicionais sem perturbações na estrutura existente. Embora estes requisitos pudessem ser satisfeitos por vários formatos, a utilização de XML é suficientemente simples e flexível para a tornar a mais atractiva: é possível manter a mesma estrutura e enriquecê-la progressivamente, à medida que são aplicados novos algoritmos ao material de trabalho (e.g., marcação de estruturas retóricas).

Além do requisito básico de utilizar XML, optou-se pela escolha de um formato normalizado, potenciando a reutilização. Foi escolhido o formato TEI[1] [4]. A escolha foi em parte motivada pela prévia utilização de corpora neste formato, mas o processo é aplicável a outros modelos com fins semelhantes, dos quais se pode realçar o XCES[2] [5].

A definição dos alinhamentos, possíveis a vários níveis, e.g. ao parágrafo, ao segmento, ou mesmo à palavra, está dependente das necessidades específicas das aplicações

[1] Text Encoding Iniciative.
[2] Corpus Encoding Standard for XML.

em estudo. Os casos de interesse, no contexto em que este trabalho se insere, contemplam desde alinhamentos ao segmento, mapeando segmentos de discurso entre duas línguas, até alinhamentos à palavra, em que são estabelecidas relações entre palavras ou grupos de palavras das duas línguas.

2.2 O Corpus JRC-Acquis

O corpus JRC-Acquis é constituído por um conjunto de textos de legislação comunitária da União Europeia, escritos entre 1950 e 2005. Esta colecção de textos legais é composta por aproximadamente 8000 documentos e cobre uma gama variada de domínios. O corpus está disponível nas 20 línguas oficiais da União (2005)[3]. Os países que passaram a integrar a União a partir de 2007 (Bulgária e Roménia), assim como outros que estão em processo de adesão (Croácia), já iniciaram a tradução da legislação, embora as traduções nas novas línguas ainda não façam parte do corpus.

Da introdução dada acima aos corpora paralelos, é clara a utilidade de um corpus desta natureza para aplicações de investigação em linguística. Se se considerar a dimensão em número de documentos e línguas disponíveis, este é um dos maiores corpora paralelos de acesso público (considerando o número de línguas): o texto base do corpus, assim como outra legislação comunitária, estão disponíveis nos servidores web da Comissão Europeia. Este material foi convertido para XML e alinhado à frase, através de relações n-para-n. Os alinhamentos são específicos para cada par de línguas alinhadas, tendo sido produzidas para todos os pares de combinações possíveis (em lugar de se utilizar uma língua pivot), embora existam algumas excepções. O corpus não foi corrigido manualmente [6,7].

3 XMLaligner

Considerando que o requisito base da aplicação era a visualização simultânea de documentos XML, foi necessário desenvolver toda a base de processamento XML da aplicação. Esta funcionalidade básica foi depois reutilizada na implementação de outras funções. Para além do requisito base, foram identificados outros aspectos de interesse, como a definição de modos de visualização – apenas parágrafos alinhados, parágrafos alinhados e seus parágrafos vizinhos, todo o documento visível – e capacidades de procura básica nos documentos actuais.

Foi ainda identificado um requisito importante em termos de interface gráfica: teria de ser facilmente utilizável por diferentes tipos de utilizadores, i.e., não poderia exigir conhecimento da estrutura da aplicação ou do formato XML do corpus.

3.1 Arquitectura

A arquitectura da aplicação possui três níveis (figura 1): apresentação e interacção, controlo e processamento do corpus. O primeiro nível corresponde à interface (gráfica)

[3] Checo, Dinamarquês, Alemão, Grego, Inglês, Espanhol (Castelhano), Estónio, Finlandês, Francês, Húngaro, Italiano, Lituano, Letão, Maltês, Holandês, Polaco, Português, Eslovaco, Esloveno e Sueco.

com o utilizador, o segundo controla o acesso aos documentos de forma alinhada e o terceiro contém os processadores XML de alinhamentos e de documentos do corpus.

As classes correspondentes aos processadores podem ser substituídas sem afectar adversamente o funcionamento geral da aplicação. Desta forma, é possível a extensão a outros formatos de corpora.

Além do requisito de apresentação dos dois conjuntos de parágrafos/documentos alinhados, a interface gráfica apresenta também a lista de documentos para o conjunto de línguas seleccionado e listas com as línguas disponíveis. É também revelada na interface alguma da estrutura dos ficheiros de alinhamentos, permitindo ver directamente a correspondência dos parágrafos, tal como é lida pelo processador de XML.

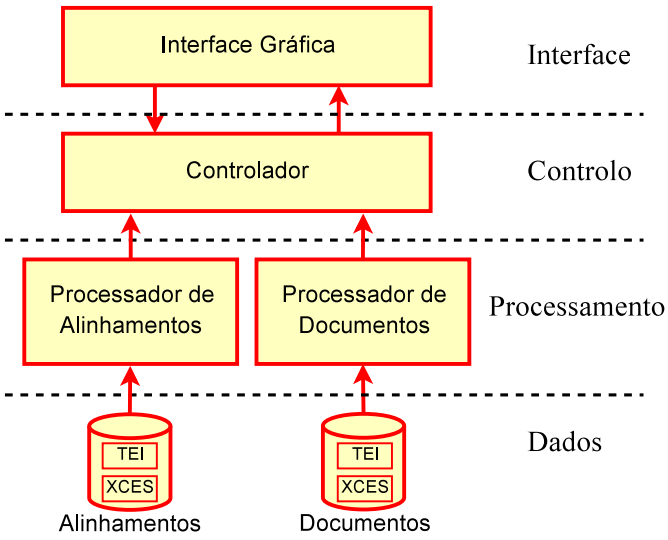

Figura 1. Arquitectura da aplicação.

As figuras 2 e 3 mostram exemplos, respectivamente, do ficheiro de alinhamentos e de um documento do corpus (ambos segundo o formato TEI).

O ficheiro de alinhamentos possui, para cada documento, uma secção `div` identificada com o código desse documento, e em que são listados o tipo de um alinhamento e os seus parágrafos-alvo. O atributo `type` indica quantos parágrafo(s) do documento na primeira língua correspondem ao(s) do documento na segunda língua; o atributo `xtargets` indica quais os números dos parágrafos referidos. A aplicação utiliza estes atributos para identificar os parágrafos que deverá mostrar, além de disponibilizar também esta informação ao utilizador.

A estrutura principal de um documento do corpus é um conjunto de parágrafos numerados, os quais são referidos nos alinhamentos. Os estádios de processamento

```
<TEI.2 id="jrc-en-pt" select="en pt">
  <teiHeader lang="en" date.created="2006-05-14">
  ...
  </teiHeader>
  <text id="jrc-en-pt." select="en pt">
    <body>
      <div type="body" n="21970A0720(01)" select="en pt">
        <p>35 paragraph links:</p>
        ...
        <link type="1-1" xtargets="21;18"/>
        <link type="2-2" xtargets="22 23;19 20"/>
        <link type="2-1" xtargets="24 25;21"/>
        ...
      </div>
    </body>
  </text>
</TEI.2>
```

Figura 2. Estrutura básica de um ficheiro de alinhamentos.

```
<TEI.2 id="jrc21970A0720_01-pt" n="21970A0720(01)" lang="pt">
  <teiHeader lang="en" date.created="2006-05-14">
  ...
  </teiHeader>
  <text id="jrc21970A0720_01-pt." lang="pt">
    <body>
      <p n="2">ACORDO COMPLEMENTAR ao ≪Acordo relativo aos
        Produtos de Relojoaria entre a Comunidade Económica
        Europeia e os seus Estados-membros e a Confederação
        Suíça≫ (1)</p>
      <p n="3">O CONSELHO DAS COMUNIDADES EUROPEIAS,</p>
      <p n="4">por um lado,</p>>
      <p n="5">O CONSELHO FEDERAL SUÍÇO,</p>
      ...
    </body>
  </text>
</TEI.2>
```

Figura 3. Estrutura básica de um documento do corpus.

permitem aceder a todos ou apenas a alguns dos parágrafos do documento, indexados pelo seu conteúdo ou número.

3.2 Construção do protótipo

O desenvolvimento inicial dos níveis de processamento começou com uma implementação SAX (Simple API for XML) do processador de ficheiros de alinhamentos. Esta implementação utilizava directamente estruturas nativas da biblioteca de suporte (`QTreeWidget` e `QTreeWidgetItem`), permitindo simultaneamente guardar os dados ao serem processados e a sua visualização directa. Foi utilizado este tipo de processamento, pois permitia um melhor controlo sobre a análise dos dados de entrada, além de possuir uma API mais directa e simples que a alternativa DOM (Document Object Model). O inconveniente da abordagem baseada em SAX era obrigar a aplicação a gerir e manter estruturas de dados para guardar a informação lida dos ficheiros XML: esta informação tinha de ser estruturada e a sua hierarquia facilmente navegável, algo que era necessário implementar explicitamente na solução SAX.

Para avaliar as capacidades do DOM, realizou-se directamente em DOM a primeira versão do processador dos documentos do corpus, em lugar de se produzir uma versão SAX, como tinha acontecido no caso dos alinhamentos. Algumas das vantagens encontradas no DOM para o caso em estudo foram:

- Os dados lidos serem automaticamente organizados numa estrutura em árvore;
- Classes DOM gerirem a criação da árvore em memória e disponibilizarem uma API para a consultar e alterar, não sendo necessário que a aplicação implemente as suas próprias estruturas para fornecer comportamento equivalente;
- As possibilidades de expansão futuras para a edição de documentos, além de consulta, sendo que o DOM permite a edição da árvore em memória, assim como a sua escrita directa para um ficheiro.

Apesar dos aspectos positivos, é, contudo, de referir que o facto de ser necessário construir e manter toda a árvore em memória pode ser uma desvantagem: ficheiros como os de alinhamentos, com um elevado número de elementos (normalmente cerca de 300.000 linhas), podem causar problemas de ocupação de memória; no caso do SAX, tem-se um maior controlo do que guardar em memória ou simplesmente descartar, o que o torna potencialmente mais eficiente, em termos de tamanho dos dados de trabalho. Para a aplicação XMLaligner, as vantagens do DOM compensam as desvantagens, pelo que todo o processamento é baseado nesta API, tendo sido reimplementada a versão inicial (SAX).

A aplicação permite ter carregado um ficheiro de alinhamentos de cada vez, podendo ser carregado outro em qualquer altura. Os documentos do corpus são carregados em memória apenas quando são consultados pelo utilizador, e descarregados assim que é escolhido um novo par. Mesmo assim, uma instância habitual desta aplicação consome cerca de 220MB de memória, o que seria expectável, tendo em conta que todos os ficheiros XML são processados com DOM e mantidos inteiramente em memória.

A pesquisa, que funciona apenas dentro do par de documentos carregado actualmente, permite a procura utilizando múltiplas palavras ou expressões regulares.

Face a outras aplicações semelhantes que correm online a partir do browser, as vantagens desta aplicação/protótipo são o trabalhar directamente sobre os dados, e permitir um nível de interactividade que seria mais difícil de obter numa aplicação baseada na web.

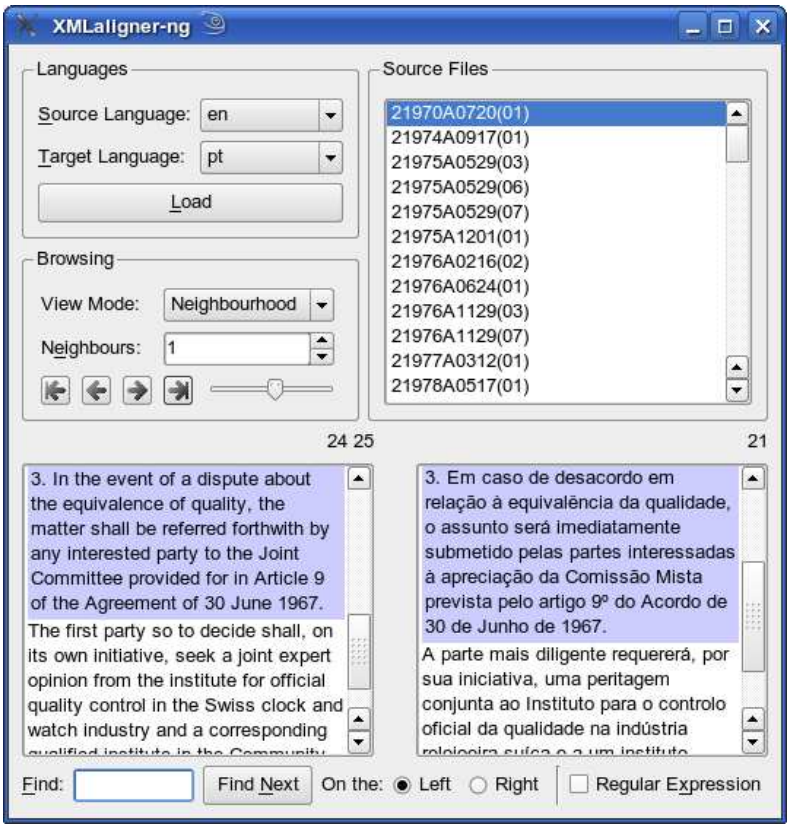

Figura 4. Protótipo apresentando alinhamentos entre as versões inglesa e portuguesa de um documento do corpus.

Para a realização da aplicação foi escolhida a biblioteca Qt 4 [8]. Este pacote providencia um conjunto extenso de componentes que permite o rápido desenvolvimento de aplicações gráficas multi-plataforma (Linux/Unix, Windows, MacOS X), que são utilizados na interface gráfica, para controlo da aplicação e apresentação de dados ao utilizador. O módulo `QtXml` contém implementações em C++ de SAX 2 e DOM (nível 2), usadas como base para os vários processadores, facilitando o seu desenvolvimento.

4 Conclusões

4.1 Protótipo

O protótipo, no seu estado actual, permite toda a navegação gráfica no corpus, incluindo escolha de alinhamentos e de documentos, procura e modos de visualização. Apesar de ainda ser apenas um protótipo, com a funcionalidade actual, já permite satisfazer parcialmente alguns dos requisitos para permitir trabalho linguístico.

4.2 Evolução

Na continuação do desenvolvimento do protótipo actual, além de melhoramentos estruturais, está planeado (i) o suporte a outros formatos de corpora, além do actual TEI; e (ii) visualização gráfica da estrutura XML dos documentos e respectiva edição.

Para desenvolvimentos futuros, além da funcionalidade presente no protótipo actual, consideram-se várias opções. A primeira é a implementação de alinhamentos a múltiplos níveis (e.g., parágrafos, segmentos e palavras), quando os corpora os possuirem, e alinhamentos de marcas, e.g., categorizações. A segunda direcção de evolução é a de providenciar métodos de busca mais sofisticados (e.g., categorias, buscas combinadas e concordâncias): os actuais métodos apenas permitem pesquisas num dos documentos de cada vez e apenas numa das línguas de cada vez. Finalmente, seria interessante considerar a ligação da aplicação de visualização a algoritmos para modificação do corpus e produção da correspondente informação de alinhamento.

Agradecimentos

Este trabalho foi parcialmente financiado pelo Projecto NLE-GRID: Natural Language Engineering on a Computational Grid (POSC/PLP/60663/2004).

Referências

1. ISO (International Organization for Standardization): ISO 8879:1986 – Information processing – Text and office systems – Standard Generalized Markup Language (SGML) (1986)
2. World Wide Web Consortium: Extensible Markup Language (XML) (2006) `http://www.w3.org/XML/`.
3. Marcu, D.: The Rhetorical Parsing of Unrestricted Texts: A Surface-Based Approach. Computational Linguistics **26**(3) (2000) 395–448
4. TEI Consortium: TEI – The Text Encoding Initiative (2006) `http://www.tei-c.org/`.
5. XCES: Corpus Encoding Standard for XML (2006) `http://www.xml-ces.org/`.
6. European Commission Joint Research Centre (JRC): JRC-ACQUIS Multilingual Parallel Corpus (2006) `http://wt.jrc.it/lt/Acquis/`.
7. Pouliquen, B., Steinberger, R.: The Acquis Communautaire Corpus. In: JRC Enlargement and Integration Workshop: “Exploiting parallel corpora in up to 20 languages”, Arona, Italy, European Commission Joint Research Centre (JRC) (2005)
8. Trolltech A.S.A.: Qt – Cross-Platform C++ Development Framework (2006) `http://www.trolltech.com/products/qt/`.