Avaliação de Alinhadores à Frase

Alberto Manuel Simões, José João Almeida

albie@alfarrabio.di.uminho.pt, jj@di.uminho.pt.

Linguateca — Pólo de Braga

Projecto Natura — DIUM

Alinhadores

- à frase;
  - memórias de tradução;
  - corpora paralelos;
- ao segmento...
- à palavra (ou termo);
  - dicionários de tradução;
  - CLIR - Cross Language Information Retrieval;
- ao caracter...?

Alinhadores à frase

Habitualmente classificados como:

- estatísticos;
  usam apenas informação presente nos corpora: ocorrências e co-ocorrências de palavras, n-grams, cognates...
- linguísticos;
  usam dicionários bilingues, analisadores morfológicos, etc, para definir com maior exactidão pontos de relacionamento;
- híbridos;

Alinhamento à frase 2

Quais os componentes de um alinhador à frase?

- exemplo: TRADOS WinAlign: (híbrido)
  1. segmentador;
  2. pré-alinhador;
  3. editor interactivo;
- exemplo: Vanilla Aligner: (estatístico)
  1. alinhador em texto segmentado (e tokenizado);
- exemplo: easy-align: (híbrido)
  1. alinhador em texto segmentado;

Arquitectura sugerida

1. segmentador:

$$texto \longrightarrow frase^{n} \qquad (bis)$$

2. alinhador:

$$\begin{aligned} frase^{n} \times frase^{m} &\longrightarrow (frase \times frase)^{\star} \\ &\longrightarrow (frase^{p} \times frase^{q})^{\star} \end{aligned}$$

3. editor interactivo:

Quais os componentes que queremos avaliar?

Segmentador 1

- *normalmente* independente do processo de alinhamento;
- útil para outras ferramentas;
  - disponibilização de corpora;
  - memórias de tradução / tradução automática;
- a ter em consideração:
  - noção de *frase* discutível;
  - tratamento de ruído...

Deverá ser avaliado independentemente?

Noção de Frase

consenso da noção de frase imprescindível;
http://acdc.linguateca.pt/treebank/CriteriosSeparacao.
como ponto de partida?

"*A velhota puxou os óculos para baixo e, por cima deles, olhou o quarto em volta; tornou a puxá-los para cima e olhou através deles.*"

"*Veio o chefe do correio, velho e pobre, que tinha conhecido melhores dias; o corregedor e a mulher, porque, entre outras coisas desnecessárias, havia ali um corregedor; o juiz de paz; a viúva Douglas, loira [...]*"

Ruído

o segmentador deve ser robusto:

"*O jogo iniciou de forma violenta... P2-5. P5-4. K9-2. ...*"

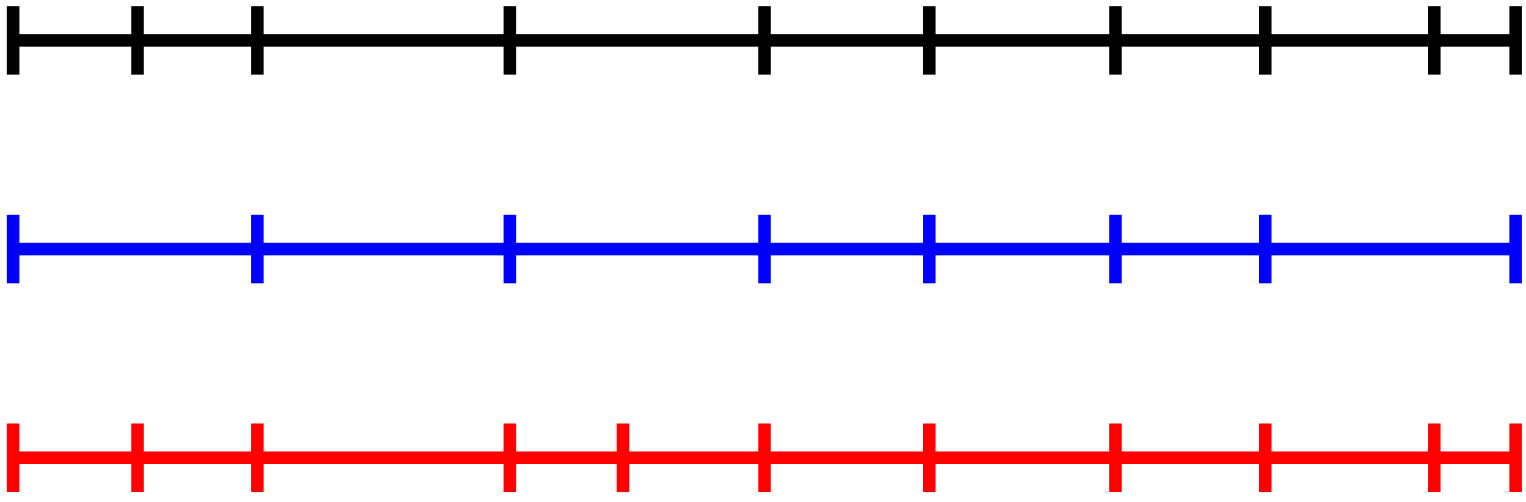
"*Para mais informações consulte http://www.linguateca.pt*"

Segmentador 2

Como se fará a sua avaliação?

1. bateria de testes, reutilizando:
   - recursos linguateca: floresta sinta(c)tica, CETEMPúblico;
   - outros que venham a surgir;
2. possíveis formas de comparação:
   - rígida entre segmentos produzidos;
   - segmentos produzidos devem ser múltiplos dos de teste;
   - pontos de corte com pesos distintos;

Comparação



Alinhador 1

O que deve ser avaliado?

- robustez (grandes quantidades);
- tipos de alinhamento:
  - $1 - 1$ (linear);
  - $n - 0, n \geq 1$ (remoção);
  - $0 - n, n \geq 1$ (inserção);
  - $n - m, n, m > 1, n = m$ (agregação);
  - $n - m, n, m > 1, n > m$ (redução);
  - $n - m, n, m > 1, n < m$ (expansão);

Alinhador 2

Como deve ser avaliado?

- análise de alinhamento real;
- quantificar a ocorrência de cada caso em situações *normais*;
- construção de bateria de testes;
  - corpus do Parlamento Europeu;
  - legendas/ficheiros de i18n;
- desenvolver analisador de resultados;

Exemplo 1 - 2

Porém, Tom não esperou pelo resto, e, no momento em que ia a sair da porta, prometeu:	But Tom did not wait for the rest. As he went out at the door he said:
- Hei-de te dar uma chibatada por conta disso, Siddy.	Siddy, I'll lick you for that!

Exemplo 3 - 1

It was not dark, yet.	Ainda não estava escuro.
Presently Tom checked his whis-tle. A stranger was be-fore him – a boy, a shade larger than himself	Pouco depois, ao ver na sua frente um rapaz mais alto do que ele, Tom moderou o tom do assobio.

Exemplo 0 - 1

	- Ora !
Never you mind what she said, Jim.	Não te importes com o que ela diz.

Editor

- avaliação não automática;
- consiste em avaliar interface e funcionalidades;

A sua avaliação não é premente...

Para que possamos avaliar...

- lista de segmentadores/alinhadores interessados em participar;
- comissão “científica” para:
  - discutir objectivos da avaliação;
  - definir módulos a serem avaliados;
  - definir formatos e casos de teste;
- publicar proposta:

  http://linguateca.di.uminho.pt/avalinha.html

- iniciar discussão de alinhamento à palavra?