CAPITULO 18

AVALIAÇÃO DE ALINHADORES

ALBERTO SIMÕES E JOSÉ JOÃO ALMEIDA

1. INTRODUÇÃO

Uma das tarefas mais importantes na preparação de corpora paralelos para o seu processamento computacional é o alinhamento: constituição de relacionamentos entre blocos destes corpora. Este alinhamento pode ser realizado usando blocos de granularidades diferentes. Habitualmente, o alinhamento é realizado ao nível da frase, do segmento ou da palavra. Por exemplo, o alinhamento ao nível da frase consiste na associação a cada frase de determinado texto da sua respectiva tradução.

Neste capítulo apresenta-se uma metodologia para a avaliação de alinhadores à frase e para alinhadores à palavra.

2. ALINHAMENTO À FRASE

Por alinhador à frase entende-se uma ferramenta que define relacionamentos entre frases de corpora paralelos. Os alinhadores à frase podem usar diferentes métodos para realizar esta tarefa. De acordo com Abaitua (2002), podem ser classificados em:

- **métodos estatísticos**: usam essencialmente os tamanhos das frases, frequências de palavras, frequências de co-relação de palavras e de elementos não textuais (Gale & Church, 1991);
- **métodos linguísticos**: baseiam o alinhamento apenas em informação linguística e léxica, nomeadamente em dicionários bilingues;
- **métodos híbridos**: combinam a informação estatística e linguística.

O processo de alinhamento pode ser dividido em duas etapas:

- detecção das frases de cada um dos corpora;
- inter-relacionamento das frases previamente detectadas.

Embora este processo seja constituído por duas etapas, é habitual que as ferramentas de alinhamento incluam apenas a segunda. Este facto obriga-nos a definir diferentes métodos para a avaliação de acordo com os seus constituintes:

- avaliar apenas a segunda etapa leva-nos a que as ferramentas com detecção de frases fiquem desfavorecidas, porque em princípio não estarão habilitadas a receber texto já segmentado:
  - No projecto PESA (Caseli & Nunes, 2003), foi realizada a avaliação de cinco alinhadores (que usam métodos variados, que se encaixam nas várias categorias apresentadas). Estes cinco alinhadores usam texto pré-segmentado, pelo que esta abordagem mantém as várias ferramentas em igualdade.
  - Também no projecto ARCADE (Véronis & Langlais, 2000) se seleccionou texto pré segmentado para a avaliação do alinhamento à frase, já que, como dizem os autores, a segmentação de texto é demasiado controversa. Deste modo, considera-se a segmentação como um dado adquirido.
- realizar uma avaliação separada para cada uma das etapas obriga à definição de uma metodologia de avaliação para a segmentação e uma outra para o alinhamento à frase, considerando o texto já segmentado. No entanto, e assim como no caso anterior, esta abordagem levará a que as

ferramentas que realizam segmentação do texto fiquem desfavorecidas já que também serão avaliadas usando texto já segmentado. Por outro lado, nem sempre será possível avaliar o resultado do segmentador da ferramenta, porque esta pode não permitir aceder ao resultado intermédio (pós-segmentador e pré-alinhador);

- proceder à avaliação do processo completo e avaliação apenas da segunda etapa: neste caso, as ferramentas serão avaliadas para aquilo que foram desenhadas, obrigando, no entanto, à definição de metodologias diferentes de acordo com a incorporação ou não de um segmentador na ferramenta de alinhamento.

Esta última hipótese parece a mais apropriada, embora leve a que não se possam comparar os alinhamentos realizados por ferramentas de categorias diferentes.

Convém ainda salientar que os alinhadores linguísticos, ou híbridos, usam recursos externos, como dicionários bilingues. Para uma avaliação rigorosa do algoritmo de alinhamento usado, seria necessário proceder à avaliação separada do algoritmo e dos recursos usados. Outra hipótese consiste em construir recursos de maneira cooperativa, para que todos os alinhadores tenham acesso a recursos da mesma qualidade.

2.1 **Avaliação baseada em corpora pré-segmentados**

Um corpus previamente segmentado corresponde a um conjunto de frases etiquetadas. É habitual que estes corpora contenham não só a marcação de segmentos, mas também a marcação de parágrafos que, por sua vez, normalmente se encontram previamente alinhados e servem de pontos de sincronização ou ancoragem nos algoritmos de alinhamento (alinhamento a vários níveis).

Cabe à ferramenta de alinhamento a definição de relacionamentos entre as frases etiquetadas nos corpora.

A avaliação destas ferramentas é baseada na comparação de um alinhamento óptimo[1] com o resultado de cada uma das ferramentas. Para que se possa usar este método de avaliação é necessária a construção de uma bateria de testes.

No projecto ARCADE (Véronis & Langlais, 2000), o alinhamento óptimo foi obtido correndo um alinhador automático e realizando uma revisão manual por dois avaliadores. Sempre que estes avaliadores discordavam num resultado de alinhamento, este era discutido até haver consenso.

Contudo, em Santos (1998) foi discutido o caso de alinhamento em corpora multilingues, em que é possível não existir mesmo uma segmentação apropriada para alinhar textos em três línguas diferentes, o que levanta o problema de poder não haver consenso em alguns casos.

2.1.1 *Definição do formato dos testes*

Para a construção desta bateria de testes, é importante definir o formato de marcação a usar. Este formato deve permitir diferenciar frases dos parágrafos, e deve permitir o relacionamento entre os parágrafos alinhados nos dois corpora. Para facilitar a avaliação é importante que a marcação inclua uma identificação de cada uma das frases.

Outro ponto a ter em consideração é que, por vezes, um alinhador pode realizar um alinhamento errado que comprometa o resultado do alinhamento do texto completo[2]. Assim, é aconselhável a criação de uma bateria grande de pares de ficheiros independentes.

---

[1] *Alinhamento realizado manualmente, por exemplo.*

[2] *Na verdade, é provável que só volte ao alinhamento correcto assim que encontrar um novo ponto de sincronização.*

Sugerimos o seguinte DTD:

```
<!ELEMENT ficheiro (paragrafo)*>
<!ATTLIST ficheiro id CDATA>
<!ELEMENT paragrafo (frase)*>
<!ELEMENT frase (#PCDATA)>
<!ATTLIST frase id CDATA>
```

Propõe-se o uso de XML para a anotação dos casos de teste (que poderão ser convertidos para os formatos usados por cada alinhador). O formato de definição de alinhamento poderá ser descrito em formatos mais flexíveis como o XCES – Corpus Encoding Standard for XML[1] ou mesmo o TEI – Text Encoding Initiative[2], embora este último tenha vindo a cair em desuso em favor do primeiro.

Johansson *et al.* (1996) foram dos primeiros a definir um alinhamento de corpora paralelos seguindo esta filosofia, no English-Norwegian Parallel Corpus (ENPC) (Oksefjell, 1999). Para um avaliação de alinhamento neste formato, veja-se Santos & Oksefjell (2000).

No projecto PESA (Caseli & Nunes, 2003), também foi utilizado um formato isomórfico para armazenar os testes, nomeadamente usando etiquetas de nomes diferentes.

2.1.2 *Definição dos casos de testes*

A definição dos casos de teste a serem usados deve tentar simular o esperado em corpora reais. O alinhamento à frase é, em cerca de 90 por cento dos casos, o relacionamento unívoco entre uma frase de cada um dos corpora. Os restantes 10 por cento correspondem a relacionamentos mais complexos, nomeadamente:

- adição ou remoção de frases — relacionamentos de uma ou mais frases para nenhuma;
- divisão ou junção de frases — relacionamentos de duas ou mais frases para uma;
- aglutinação de frases — relacionamentos de duas ou mais frases para outras duas ou mais frases, em que a semântica não permite que sejam separadas em dois relacionamentos.

A distribuição das frequências das diferentes formas de alinhamento pode ser extraída automaticamente de corpora alinhados manualmente. Por exemplo, o COMPARA (Frankenberg-Garcia & Santos, 2002) inclui uma página de estatísticas, que inclui a distribuição das formas de alinhamento[3]. A título demonstrativo, na versão 6.7.1, o COMPARA apresenta 90% de alinhamentos de 1-1 unidades, 4,3% de alinhamento 2-1, 3,7% de alinhamento 1-2, 0,45% de alinhamentos 1-0.

No PESA foram utilizados cinco corpora de estilos diferentes, alinhados manualmente. Embora garanta que as quantidades dos vários tipos de alinhamento surjam com a probabilidade esperada, não permite a fácil divisão em pequenos ficheiros de teste.

Também no projecto ARCADE, o corpus de teste inclui textos de diferentes fontes, desde literatura, textos jurídicos e manuais técnicos.

2.1.3 *Definição do modelo de alinhamento*

Tendo como base estes dois pontos, procede-se à criação da bateria de testes, usando textos reais, escolhidos manualmente (ou automaticamente, usando um corpus alinhado manualmente). Dever-se-á proceder à estatística dos casos encontrados, para que o conjunto de testes esteja equilibrado.

---

[1] *http://www.cs.vassar.edu/XCES/*

[2] *http://www.tei-c.org/*

[3] *http://www.linguateca.pt/COMPARA/Conteudo.html*

Segue-se um pequeno exemplo de um par de textos anotados de acordo com o DTD proposto[1].

```
<ficheiro id="2">
  <paragrafo>
   <frase id="1">
    Só quando passou por eles, levando um enorme donut num saco,
    conseguiu apanhar no ar algumas palavras do que eles estavam a
    dizer.
   </frase>
   <frase id="2">
    - Os Potters, sim, foi o que ouvi dizer.
   </frase>
   <frase id="3">
    - Sim, o filho deles, Harry.
   </frase>
   <frase id="4">
    O medo apoderou-se dele .
   </frase>
  </paragrafo>
 </ficheiro>

<ficheiro id="2">
  <paragrafo>
   <frase id="1">
    It was on his way back past them, clutching a large doughnut in a
    bag, that he caught a few words of what they were saying.
   </frase>
   <frase id="2">
    «The Potters, that's right, that's what I heard yes, their son, Harry»
   </frase>
   <frase id="3">
    Mr. Dursley stopped dead.
   </frase>
   <frase id="4">
    Fear flooded him.
   </frase>
  </paragrafo>
 </ficheiro>
```

[1] *Exemplo transcrito do livro* Harry Potter and the Philosopher's Stone *de J.K. Rowling, traduzido por Isabel Fraga como* Harry Potter e a Pedra Filosofal.

O passo seguinte é a definição do modelo de resultado. Uma vez que cada frase está univocamente identificada, pode utilizar-se o seu identificador no modelo.

Voltando a utilizar o exemplo anterior, poder-se-ia considerar que o alinhamento correcto (ou, pelo menos, o mais adequado) podia ser indicado do seguinte modo:

ING - PORT

1 - 1

2 - 2

3 - 2

0 - 3

4 - 4

Repare-se que, no caso da segunda e terceira frases do corpus de origem que alinham com a mesma frase (segunda) do corpus de destino, são criadas duas entradas no ficheiro de alinhamento. Da mesma forma, a terceira frase do corpus de destino, que não tem alinhamento possível, é alinhada com a pseudo frase número zero. Um resultado com este formato pode ser comparável de maneira simples com o modelo esperado.

Contudo, formas alternativas de codificar a situação acima descrita, são:

1 - 1	1 - 1	1 – 1
2,3 - 2	2,3 - 2,3	2,3 - 2
0 - 3	4 - 4	4 - 3,4
4 - 4		

Note-se que em nenhum dos casos se entra em conta com o facto de haver um alinhamento parcial (no sentido de que as correspondências podem não ser totais). Poder-se-ia imaginar um modelo mais sofisticado de alinhamento, que indicaria, além da relação entre os números das frases, se era correspondência total, inclusão, ou sobreposição parcial.

2.1.4 *Medidas de qualidade*

Para descrever a qualidade dos resultados obtidos, uma série de medidas podem ser usadas. A precisão e a cobertura do alinhamento poderão ser calculadas usando as fórmulas habituais para este tipo de medidas, definindo como alinhamentos de referência o primeiro modelo anterior:

$$\textit{cobertura} = \frac{\text{Número de alinhamentos correctos}}{\text{Número de alinhamentos de referência}}$$

$$\textit{precisão} = \frac{\text{Número de alinhamentos correctos}}{\text{Número de alinhamentos propostos}}$$

Embora a distribuição da quantidade das diferentes formas de alinhamento dê mais peso aos alinhamentos simples (1–1), na falta de exemplos suficientes para manter a distribuição poder-se-ão associar pesos diferentes a cada um destes tipos de alinhamento, criando assim uma distribuição artificial. Estas medidas podem, além disso, ser pesadas em função do número de palavras envolvidas, ou da diferença de número de palavras em cada par, já que é mais difícil alinhar frases com tamanhos diferentes.

2.2 **Avaliação baseada em corpora não segmentados**

A avaliação de alinhadores à frase baseada em corpora não segmentados levanta mais problemas do que a apresentada anteriormente:

- em que formato fornecer os corpora: texto, HTML ou RTF são vários dos formatos suportados por alguns destes alinhadores;
- uma vez que as frases vão ser definidas pelos alinhadores, não é possível associar um identificador a cada uma, para a definição de um modelo de avaliação;
- enquanto, na avaliação baseada em corpora segmentados, a definição de frases, mesmo que não consensual, não irá ter repercussões no alinhamento, o mesmo não é verdade na avaliação baseada em corpora não segmentados. Portanto, a escolha de exemplos para a bateria de testes em que a segmentação levanta falta de concordância entre os membros da avaliação conjunta deve ser evitada.

A solução proposta passa pela definição de um modelo de segmentação e de alinhamento no qual se tentará encaixar a segmentação efectuada pelo alinhador a avaliar. A criação deste modelo passará pela segmentação e pelo alinhamento da bateria de testes. Esta segmentação e alinhamento deverão ser validados manualmente.

O processo de comparação dos resultados a avaliar passará por encaixar os segmentos criados pela ferramenta nos segmentos do modelo.

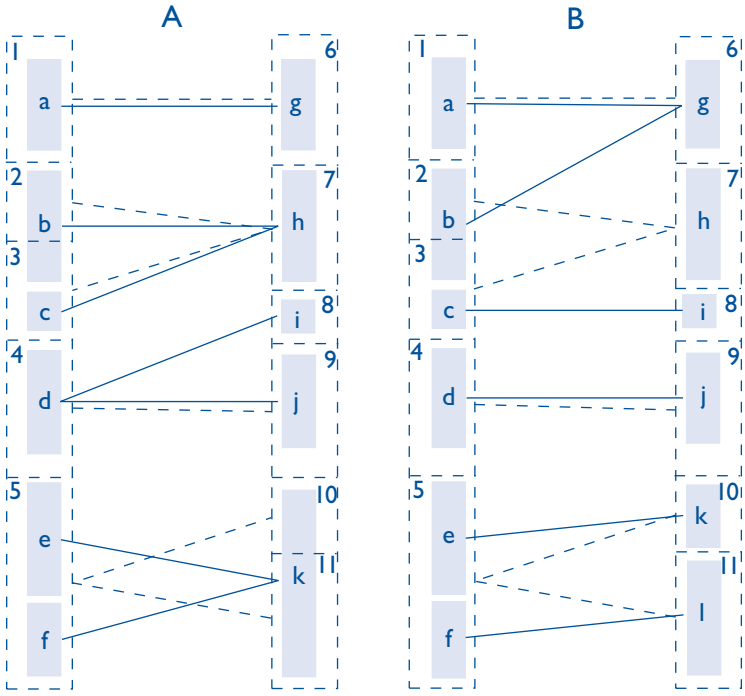

Figura 18-1

Comparação de dois segmentadores-alinhadores fictícios

Na Figura 18-1, as linhas tracejadas correspondem à segmentação e ao alinhamento do modelo, no qual se tentou encaixar o alinhamento de duas ferramentas diferentes (A e B).

A ferramenta A tem um alinhamento bastante aceitável. Embora em certos casos a segmentação não coincida com a segmentação do modelo, o seu encaixe no alinhamento é perfeito (só o segmento **i** foi incorrectamente alinhado).

Embora a ferramenta B tenha segmentação também aceitável, o seu alinhamento já não o é. Por exemplo, é complicado pensar como será avaliado o alinhamento **e–k** e **f–l**, já que divide um bloco do modelo. Mais simples é verificar que o alinhamento **a,b–g** está incorrecto.

A avaliação do alinhamento à frase irá passar pela análise dos possíveis encaixes, e classificação (atribuição de pesos) dos diferentes tipos de falhas de alinhamento e de segmentação.

2.3 **Construção e obtenção dos casos de teste**

Quer no caso de utilização de corpora pré-segmentados ou não, uma das tarefas mais importantes antes da realização da avaliação dos alinhadores, é a construção da bateria de testes.

Como foi referido, a quantidade de diferentes tipos de alinhamento deve tentar simular a probabilidade de serem encontrados em corpora reais. Deste modo, a primeira tarefa (quando possível) será a análise de ocorrências dos vários tipos de alinhamento num corpus alinhado manualmente.

Obtidas estas estatísticas, procede-se à procura de casos de teste para serem incluídos. Este processo será, sem dúvida, o mais moroso.

Existindo corpora grandes segmentados e alinhados à mão, poderão ser utilizados para, de uma maneira automática, extrair casos de teste. Por exemplo, no caso do português/inglês, poderia ser usado o COMPARA. No entanto, como a licença dos textos usados no COMPARA não nos permite a sua distribuição, teria de ser realizada uma extracção de parágrafos. Dado o formato interno do COMPARA e as estatísticas de tipos de alinhamento presentes na página já mencionada, será possível recolher casos de teste específicos para cada um dos tipos de alinhamento, bem como a extracção de um conjunto de frases de contexto.

Na falta deste tipo de recurso, a criação da bateria de testes irá passar pelo alinhamento automático de corpora, e análise manual do alinhamento.

Em qualquer um dos casos, é sempre possível forjar um tipo de alinhamento mais difícil de encontrar, editando directamente o corpus, já que não há necessidade de que os casos de teste sejam reais.

Outro dos pontos importantes durante a construção dos casos de teste é a quantidade de frases existentes para o alinhamento. Deste modo, é importante que se construa um teste específico para o teste de robustez da ferramenta.

3. ALINHAMENTO À PALAVRA

O termo «alinhamento à palavra» é ambíguo, dadas as duas interpretações apresentadas na bibliografia da área. A única semelhança consiste no facto de relacionarem palavras. Uma das abordagens (Dagan *et al.*, 1993) defende que o alinhamento à palavra consiste em, para cada frase e respectiva tradução (a que chamaremos unidade de tradução), associar a cada palavra de uma das línguas uma ou mais (ou mesmo nenhuma) palavra da outra língua. A outra abordagem (Hiemstra, 1996) defende que se pretende obter no final apenas um relacionamento entre palavras de uma língua e as respectivas traduções usadas no corpus. Este relacionamento associa a cada palavra e respectiva tradução uma medida probabilística de corresponder a uma tradução válida.

Obviamente, faz sentido realizar a avaliação em ambas as abordagens, ou tipos de alinhamento à palavra.

3.1 **Avaliação de alinhamento palavra a palavra em textos paralelos**

Sobre a primeira abordagem, existe uma descrição muito pormenorizada em Mihalcea & Pedersen (2003), que resumimos aqui, notando que a forma sugerida muito se assemelha à avaliação proposta para o alinhamento à frase.

Tomemos como exemplo a segunda frase do extracto anterior para o alinhamento à frase, que de seguida se repete:

<frase id="2">
- Os Potters , sim , foi o que ouvi dizer .
</frase>

<frase id="2">
" The Potters , that's right , that's what I heard yes , their son , Harry "
</frase>

A proposta de resultado para o alinhamento à palavra destes extractos é composta por três colunas: o identificador da frase que estamos a avaliar, e um identificador de palavra de cada frase da unidade de tradução[1]:

2 2 2 (os - the)
2 3 3 (Potters - Potters)
2 5 5 (sim - that's)
2 5 6 (sim - right)
2 7 8 (foi - that's)
2 8 9 (o - what)
2 9 9 (que - what)
...
2 0 14 (null - their)
2 0 15 (null - son)
2 0 17 (null - harry)

Neste formato são repetidos os identificadores das palavras que se alinham com mais do que uma palavra, e as palavras que não têm palavra respectiva são alinhadas com a palavra fictícia, identificada por 0.

3.2 **Avaliação de dicionários probabilísticos de tradução**

Nesta secção defendemos uma metodologia para a avaliação do resultado do segundo tipo de alinhamento, a que iremos chamar de dicionários probabilísticos de tradução.

Um dicionário probabilístico de tradução consiste num mapeamento de palavras de uma língua com as suas possíveis traduções dado um corpus ou conjunto de corpora, bem como uma medida probabilística da correcção da tradução.

---

[1] *Para simplificar o exemplo, vamos omitir o alinhamento dos sinais de pontuação. Também apresentamos uma coluna adicional com as palavras em causa, para facilitar a leitura.*

coruja	owl	97%
	vacuum	2%
	forward	1%
enorme	large	42%
	huge	23%
	enormous	13%
	(null)	11%
	deep	4%

Figura 18-2
Duas entradas de um dicionário probabilístico de tradução fictício

Este dicionário indica que a palavra *coruja* foi a tradução (ou terá sido provavelmente a tradução) de *owl* em 97 por cento das vezes, de *vacuum* em dois por cento das vezes e de *forward* em um por cento das vezes.[1] O mesmo se interpreta para a entrada de *enorme*. A única diferença corresponde à palavra fictícia *(null)*, que corresponde ao alinhamento com nenhuma palavra. Note-se que, embora os exemplos aqui apresentados correspondam a relações de uma palavra com outra (ou nenhuma, no caso do *null*), é possível e natural que existam entradas de várias palavras (por exemplo, *comerei - will eat*).

Um algoritmo de cálculo destes dicionários, envolvendo o português como uma das línguas, encontra-se descrito em Simões (2004).

Mais uma vez, a avaliação irá passar pela construção de um modelo que será comparado com os resultados das ferramentas a avaliar. Este modelo, construído manualmente, deverá apresentar variedade não só no tipo de categorias gramaticais incluídas (verbos, substantivos, preposições, pronomes), mas também no tipo do alinhamento esperado (alinhamento para nenhuma, uma ou mais palavras).

Este modelo, para um conjunto de entradas seleccionadas, pode ser obtido por escolha manual em corpora bilingues, para o caso de formas suficientemente frequentes.

Uma primeira abordagem pode limitar-se a comparar as traduções por cada palavra, ordenadas por probabilidade, e considerar que o alinhamento está de acordo com o modelo, se as traduções forem as mesmas, e pela mesma ordem.

No entanto, e dado que estamos a analisar dicionários probabilísticos, não devemos descurar as probabilidades, e portanto, a comparação deve ser feita tendo em consideração probabilidades dentro de uma gama de valores semelhantes.

Por outro lado, e para obter um conjunto de teste de maior variabilidade, é possível também comparar outro conjunto de dados com as entradas de um dicionário bilingue tradicional (que não contém, evidentemente, probabilidades), embora seja de considerar a hipótese de este não reflectir o género textual em questão. Veja-se, para este tipo de abordagem, Karlgren & Sahlgren (2005).

4. CONCLUSÃO

A definição de uma metodologia para a avaliação de alinhadores à frase é possível. No entanto, é muito mais simples a avaliação apenas da tarefa de alinhamento, do que a avaliação do par segmentação/ /alinhamento.

[1] *Note-se que, sendo o resultado de um modelo probabilístico, o conteúdo destas entradas foi calculado automaticamente com base num presumível alinhamento palavra-a-palavra subjacente, e pode não corresponder a casos reais de a palavra* coruja *ter sido uma tradução de* vacuum.

A principal diferença reside no facto de que a avaliação apenas da tarefa de alinhamento cinge-se a uma simples comparação de resultados, enquanto a avaliação do par segmentação/alinhamento obriga à definição de um algoritmo complexo de encaixe de segmentos.

Os projectos de avaliação de alinhadores (como o PESA ou o ARCADE) têm vindo a restringir-se apenas à avaliação do alinhamento, não só pela sua maior simplicidade como também porque não existem muitos alinhadores que incluam a segmentação. Dado que a segmentação é útil em vários campos do processamento da linguagem natural, tem vindo a ser desenvolvida de forma separada das restantes aplicações. Este facto leva-nos a concluir que a avaliação de segmentadores será, também, interessante, se for possível definir um contradomínio de especificações consensual, suficientemente limitado.

Quanto à avaliação de alinhadores ao nível da palavra, esta já tem vindo a ser realizada, embora não para o português, ao nível de encontros internacionais de avaliação, como por exemplo como satélite do HLT/NAACL 2003 (Mihalcea & Pedersen, 2003). No entanto, a avaliação de dicionários probabilísticos de tradução tem sido esquecida. Dada a diferença entre os dois tipos de ferramentas envolvidos nestas duas abordagens, torna-se imperativa a distinção, e cremos que deve passar a considerar-se – e avaliar-se – os dicionários probabilísticos de tradução, de modo independente do processo de alinhamento.