Capítulo 1.
Introdução ao modelo de avaliação conjunta

Diana Santos

Este capítulo pretende apresentar o paradigma da avaliação conjunta, uma metodologia de avaliação que visa promover a investigação numa tarefa através da comparação de vários sistemas com base em recursos e tarefas comuns. São esboçadas as características, os problemas, as vantagens e a história da avaliação conjunta na engenharia da linguagem.

1 Apresentação

Avaliação conjunta foi a forma como baptizámos, em português, o que em inglês tem vindo a ser chamado "evaluation contest", "evaluation campaign" ou "joint evaluation", e que é um actividade que junta várias participantes cujos sistemas são comparados ao executar uma tarefa comum.

Ainda que a palavra avaliação tenha muitas conotações, algumas delas nem sempre positivas, este novo termo pretende sublinhar o esforço "em conjunto", que permite documentação, confronto com questões novas e a necessidade de concordar num conjunto de especificações.[1]

Pretende-se, acima de tudo, estabelecer em conjunto os objectivos e os factores de sucesso de uma dada actividade, fomentando ao mesmo tempo o diálogo entre os diversos actores. Desde que a organização junte o esforço necessário, é também possível e desejável criar um recurso que permita a novos investigadores não começar do zero e que possibilite a todos os membros da comunidade a utilização e melhoria do trabalho já feito. Adicionalmente, esta forma de organização conjunta permite resolver algumas questões que, parecendo mínimas ou mesmo irrelevantes, podem dar origem a muita frustração e sobretudo falta de comparabilidade.

---

[1] A própria disciplina de avaliação tem um âmbito imenso, que não será sequer mencionado aqui. Veja-se Santos (2000a) para uma possível introdução à área da avaliação de sistemas de processamento de linguagem natural (PLN) e à problemática da avaliação em geral.

Como explicam Voorhees & Tice (2000b:206), o primeiro objectivo de uma avaliação conjunta é promover a investigação na tarefa. Um segundo objectivo (por parte da organização) é investigar se a metodologia de avaliação é apropriada, e se através dela é possível definir recursos de avaliação reutilizáveis, visto que "colecções de teste reutilizáveis, que permitem que os investigadores experimentem com ela e recebam uma resposta rápida sobre a qualidade de métodos alternativos, são fulcrais para avançar o estado da arte." (Voorhees & Tice, 2000b:207)

Em suma, no modelo de avaliação conjunta o que se pretende é multiplicar o número de beneficiários de uma dada actividade e melhorar em conjunto o estado da área, evitando reinventar a roda, aumentando o número de trocas científicas entre grupos distintos, e produzindo padrões de funcionamento que evitem (a partir dali) a possibilidade de grupos novos começarem de novo com uma autoavaliação.

Este livro, que é um dos resultados da primeira avaliação conjunta para o português, as Morfolimpíadas, relata essa experiência em pormenor. No que resta deste capítulo, o paradigma é apresentado em detalhe, esboçando-se as características, os problemas, as vantagens e a história da avaliação conjunta na engenharia da linguagem.

2 O modelo da avaliação conjunta

2.1 Modelos de avaliação anteriores

Para explicar porquê falar de um novo paradigma – novo, note-se, para a língua portuguesa, uma vez que já tem sido abundantemente aplicado para o inglês e outras línguas, como se verá na secção 6 – convém referir os dois paradigmas anteriores que presidiam ao processamento do português.

- autoavaliação: um investigador, ou um grupo reduzido, criava um sistema para ilustrar um método ou uma teoria e, quando muito, publicava resultados de avaliação segundo as suas próprias premissas, sem garantia de replicabilidade ou de verificação por outros grupos;
- modelo empresarial: uma empresa desenvolvia um sistema para um dado negócio, provavelmente com base numa autoavaliação, e a verdadeira avaliação era feita pelos utilizadores na forma de realimentação (“feedback”) a ser incorporada em novas e melhores versões do sistema, caso a empresa decidisse continuar o desenvolvimento.

Sem querer criticar demasiado estes paradigmas, é de salientar que em nenhum dos casos o conhecimento obtido era reutilizável por uma comunidade científica (no sentido lato, incorporando também desenvolvedores e testadores, e não apenas investigadores). Muitas vezes, esse conhecimento morria nas organizações (universidades ou empresas) devido à mobilidade dos investigadores ou ao encerramento dos projectos. No melhor dos casos, mantinha-se propriedade e conhecimento (por alguns chamada vantagem competitiva) de um único grupo.

2.2 Características principais

As características principais de uma avaliação conjunta são, no meu entender, as seguintes:

- um debate inevitável de forma a obter um consenso, implicando o conhecimento mútuo de diferentes pontos de vista e de diferentes actores com preocupações distintas;
- a identificação de problemas a resolver e de problemas já resolvidos;
- a clarificação e fixação de terminologia – embora não seja estritamente necessário, e poderá ser mesmo em alguns casos impossível, que todos concordem nos termos a utilizar, é desejável que se definam claramente as relações de sinonímia ou não-sinonímia no vocabulário de uma área comum;
- a obtenção, minuciosa e formal, de um conjunto de resultados sobre os quais há consenso, ou sobre os quais há uma noção clara de várias alternativas possíveis;
- a definição de um conjunto de tarefas objectivas que os sistemas devem efectuar;
- a identificação de diferenças irredutíveis, e de zonas cinzentas, ambas importantes para a documentação da área e do seu progresso;
- a impossibilidade de não citar, ou de deixar implícitas, algumas questões relevantes e que podem polarizar consideravelmente a avaliação – o que acontece, necessariamente, quando o mesmo grupo desenvolve o sistema e o método de avaliação.

Concretizando, de um ponto de vista factual, uma avaliação conjunta tem normalmente os seguintes ingredientes:

- vários sistemas participantes
- uma organização com conhecimento do assunto

- a possibilidade de produzir resultados e partilhá-los com os participantes
- a possibilidade prática de comparar esses resultados (por outras palavras, os resultados produzidos pelos vários sistemas têm de ser minimamente comensuráveis)

Do ponto de vista organizativo, é preciso:

- uma comissão de organização
- o reconhecimento da idoneidade da organização para que todos os participantes se encontrem em igualdade de circunstâncias
- uma forma prática de coligir e divulgar os resultados

E é, além disso, geralmente desejável que:

- os recursos de avaliação sejam posteriormente tornados públicos
- os participantes se juntem num ou vários encontros presenciais para que a troca de ideias e a consequente fertilização cruzada possam ter lugar da melhor maneira
- as agências de financiamento tenham em conta os resultados e a participação nestas avaliações como um grau de maturidade a prestigiar, senão mesmo como uma actividade que devem absolutamente financiar

Na maior parte dos casos, uma avaliação conjunta tem várias edições, como ilustrado na figura 1. Além disso, para ajudar a fixar os parâmetros, pelo menos nas edições iniciais, costuma-se usar um ensaio ("trial" ou "dry-run" em inglês). Isto significa que uma avaliação conjunta é sempre um acontecimento que decorre num período alargado.

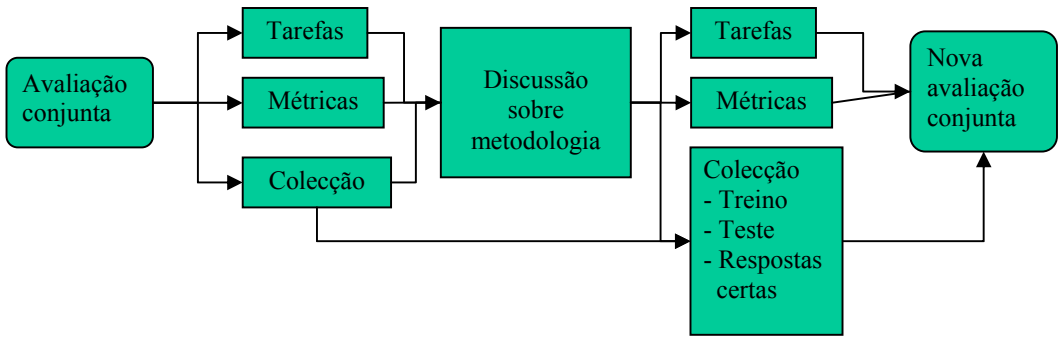

Figura 1. O desenrolar de cada edição de uma avaliação conjunta do tipo MUC

2.3 Vantagens para os sistemas participantes

Comecemos por referir as numerosas vantagens de que um grupo pode auferir ao participar numa avaliação conjunta.

Se, em primeiro lugar, se participa na fixação de um problema e no acertar na sua melhor medição, passa-se, mais tarde, a ter um fórum e um aparelho que permite separar o desenvolvimento de um sistema (a cargo do grupo) do teste de hipóteses (que é fornecido e gerido por uma organização externa).

Embora neste livro nos concentremos principalmente nas vantagens iniciais, a possibilidade de usar uma organização externa para fazer avaliações internas não é de desprezar. Clarke *et al.* (2001), por exemplo, enviam três resultados diferentes e usam o aparelho do TREC, assim como o fórum de publicações deste, como o canal ideal para medirem o progresso e o impacto de diferentes opções. Para o português, Costa (2004) usou o CLEF para comparar o desempenho do seu sistema seguindo duas abordagens consideravelmente diferentes.

Outras vantagens não desprezáveis são a existência de prazos externos, o conhecimento dos resultados de outros sistemas, e a facilidade de comparação e acesso ao estado da arte e ao resultado mínimo definido ("baseline"), a que passaremos a chamar RMD. O RMD é o resultado obtido com um algoritmo ingénuo, fácil de implementar e ao alcance de todos. Só resultados acima desse RMD são relevantes, visto que, senão, o melhor seria usar o tal sistema básico.

É também preciso sublinhar que a existência de uma comunidade de pessoas que se reconhecem num mesmo objectivo e com as quais é possível discutir ideias e observar as consequências de opções diferentes, levando a uma fertilização cruzada e a um reconhecimento público, é extremamente importante para transformar uma actividade (neste caso, o processamento do português) numa área científica respeitada e com visibilidade (e, como tal, financiada, levando-a em última análise a ser mais útil a todos os falantes).

Para que uma avaliação conjunta possa conseguir isto, é no entanto preciso que se baseie em procedimentos rigorosos e consensuais, como discutiremos na secção 4 em mais pormenor.

2.4 Algumas críticas e limitações

É preciso, no entanto, esclarecer que há críticas válidas a apontar a este modelo, e que o seu abuso pode levar a alguns inconvenientes:

1. Por um lado, há o perigo de os participantes desenvolverem os seus sistemas para participar nas competições e não para funcionar na vida real. Isto tanto pode acontecer pela concentração numa área específica (por exemplo, houve um excesso de afinação dos sistemas que concorriam no MUC sobre o assunto "actividades terroristas", por ter sido o domínio usado em duas competições seguidas), como considerando cuidadosamente as questões medidas por uma dada avaliação conjunta e desprezando as que não são objecto de comparação.

Um exemplo flagrante deste tipo de manipulação, sem qualquer contrapartida na vida real, é a inserção da resposta nula em terceiro lugar na competição Q&A-TREC[2] 2001 por Clarke *et al.* (2001). O facto de poder não haver resposta numa colecção, e como tal a resposta nula ser possível, não diz nada a respeito da qualidade do sistema de resposta a perguntas. Contudo, a forma como os sistemas eram pontuados favorecia este truque.

Nas palavras de Gaizauskas (2003), é preciso evitar que os concorrentes foquem o seu trabalho em melhorar o desempenho em décimas de grau.

2. Por outro lado, a existência dos chamados cotovelos, que indicam que se chegou a uma altura em que a maior parte dos sistemas já têm um desempenho aceitável, indicia provavelmente que se deverá mudar o cenário da avaliação radicalmente. Hirschmann (1998) comenta que, de facto, de uma edição para a próxima, deve-se sempre aumentar os desafios e evoluir na avaliação; Gaizauskas (2003) sugere que o estudo de avaliações anteriores sugere novas métricas que, em si, sugerem novos tipos de avaliação... ou seja, uma actividade de avaliação conjunta deve progredir ao longo das várias edições, aprendendo com o passado e com os erros e os sucessos dos sistemas participantes.

3. O principal perigo de uma avaliação conjunta é, contudo, no meu entender, que as medidas ou o cenário não sejam adequadas ao problema ou privilegiarem um certo tipo de funcionamento. Por exemplo, uma característica do cenário que tem sido muito criticada é a questão dos "fragmentos de 50 bytes" na avaliação de sistemas de

---

[2] Uma avaliação conjunta de resposta automática a perguntas (RAP).

resposta automática a perguntas (Q&A TREC) já mencionada. Que significado pode ter o tamanho duma resposta?

Uma objecção relacionada é a de que as avaliações conjuntas reduzem o valor de um sistema a um número, ou a uma série de números, não necessariamente facilmente interpretáveis, compreensíveis ou mesmo intuitivamente adequados. Ora as medidas têm de ser intuitivamente interpretáveis para todo o processo fazer sentido. Kilgarriff (2003) denuncia o uso indiscriminado de uma medida muito em voga, criticando a sua ubiquidade e demonstrando a sua inadequação ao problema concreto em causa.

Por outro lado, nem sempre é fácil eleger UMA medida quando há muitas variáveis que é preciso ter em conta. A este respeito, mencione-se que o TREC e o CLEF usam uma vintena de medidas.[3]

De igual modo, é muito importante compreender a junção dos resultados, quase nunca simples. Tome-se o seguinte exemplo trivial: num campeonato por equipas, um conjunto de cinco alunos com nota de 1 a uma dada disciplina e um aluno com nota de 5 mais quatro com 0, embora cumulativamente dêem o mesmo valor se a medida for "soma das notas" (5) ou "média das notas" (1), estão longe de desempenhar o mesmo papel a responder a perguntas sobre a matéria dessa disciplina!

4. De forma análoga, há o perigo da competição focar problemas não fundamentais (Kilgarriff & Palmer, 2000:9). De facto, para cada tarefa, é preciso averiguar da relevância da tarefa para a aplicação a que está subjacente, sob pena de a avaliação não ser válida. Por exemplo, é justificável definir desambiguação de sentidos como uma tarefa separada dentro do processamento de linguagem natural? Wilks (2000) critica a própria pertinência de avaliar separadamente esta tarefa, como é feito no Senseval.

5. Gaizauskas, no PROPOR 2003, referiu que uma avaliação conjunta pode ser enganosa no sentido de levar os participantes (ou os organizadores) a convencerem-se de que, só por efectuarem medidas, estão a fazer ciência. Para o fazer, é preciso

---

[3] Veja-se Gonzalez *et al.* (neste volume) para uma introdução detalhada às mais discutidas, assim como a secção 4.4.

conferir e confirmar se as diferenças entre participantes são estatisticamente relevantes. Veja-se a este propósito, Zobel (1998) sobre a relação entre os valores de desempenho e a significância estatística, em que ele demonstra, usando os resultados dos TRECs 3, 4 e 5, que algumas suposições geralmente aceites não são correctas.[4]

Ou seja, é preciso ter extremo cuidado na interpretação dos resultados e no planeamento de toda a avaliação.

Seguem-se outras críticas enunciadas, que, no meu entender, embora válidas, não desvirtuam o modelo como aqui foi apresentado:

6. As avaliações conjuntas em geral não entram em conta com a "experiência do utilizador", nem com outras qualidades subjectivas (ou mais dificilmente mensuráveis) de um sistema, tais como qualidade científica, qualidade da documentação, forma de apoio ao utilizador ou maneira como recuperam de erros, inovação, consistência, facilidade de manutenção, legibilidade do código, consistência com as teorias linguísticas mais modernas, etc.

Isto só indica que a aferição do valor e da utilidade de um sistema não se esgota com uma avaliação conjunta: outros tipos de avaliação podem também ser necessários. Contudo, em alguns casos, pode juntar-se a uma avaliação conjunta um painel de peritos que pontuam algumas destas propriedades (como nas Morpholympics organizadas por Hausser em 1994, Hausser, 1996), assim como se pode fazer avaliações com o utilizador ou com um perfil claro de utilizadores em mente, como proposto por Paroubek & Blasband (1999) (veja-se também Aires & Aluísio, neste volume).

7. As avaliações conjuntas não são adequadas para tarefas mais complicadas. Por exemplo, Sabatier *et al.* (1997) referem que não é possível fazer uma avaliação conjunta de "sistemas de compreensão de texto". Entre os vários argumentos apresentados, é referido que há conhecimento lexical e conhecimento do mundo, há tipos de textos diferentes, tipos de domínios diferentes e tipos de aplicações diferentes, etc.

---

[4] Em particular, Zobel mostra que diferenças mínimas de desempenho podem ser estatisticamente significativas, enquanto diferenças muito maiores o podem não ser.

Na minha opinião, esta observação releva de que a área, como ele a descreve, é demasiado abrangente. Uma avaliação conjunta pressupõe sempre "uma aplicação" num dado contexto, ou seja, vários tipos de textos, um ou vários domínios fixos. O que se passa, e nisso ele pode ter razão, é que sistemas diferentes igualmente "bons" podem ser incomensuráveis, se se der o caso de que um trabalha sobre acórdãos da Procuradoria da República, outro sobre resumos de revistas médicas e outro ainda sobre anúncios em jornais. Ou seja, pode não ser possível arranjar uma tarefa igualmente neutra para todos os sistemas e ao mesmo tempo justa para os avaliar.

Há, no entanto, sempre a possibilidade de fazer a união dos tipos de texto e problemas gratos a cada grupo, e medir numa avaliação conjunta a degradação dos sistemas quando aplicados a outros ambientes.

Ao contrário da conclusão de Sabatier que a comparação de sistemas sobre um mesmo domínio tem **apenas** uma função selectiva (num dado instante), estou no entanto convencida de que a especificação, em comum, de um subconjunto de questões e respostas sobre uma mesma base permite uma reflexão sobre os problemas concretos, a forma de avaliação, e as vantagens e desvantagens de diferentes opções, que é incomparavelmente mais rica do que uma argumentação teórica sobre o um conjunto de princípios independentes da aplicação.

8. Finalmente, é preciso notar que só se pode começar a aplicar o modelo de uma avaliação conjunta quando há mais de um grupo interessado, e mais de um sistema operacional numa dada área. Antes disso, a única forma de avaliação possível é a autoavaliação.

Esta observação espelha sobretudo uma questão de prioridades. Teoricamente, a Linguateca sugeriu que os interessados na área comecem a produzir recursos para a sua futura avaliação, mesmo sem terem ainda sistemas implementados. Contudo, é natural que não haja recursos humanos para organizar tal tarefa, quando ainda há tanto que fazer em outras áreas já com sistemas a funcionar.

3. Um pouco de história a nível internacional

Existem artigos muito completos que documentam a história deste paradigma, donde apenas cito alguns aqui[5]: Hirschman (1998c) para o MUC, Voorhees (2002) para o

[5] Convém mencionar que este artigo se restringe apenas à língua escrita.

TREC, Braschler & Peters (2004) para o CLEF, Kando (2002) para o NTCIIR e Edmonds & Kilgarriff (2002) para o Senseval. Além disso, as avaliações conjuntas em funcionamento no presente têm páginas na rede que são referência fundamental (TREC, NTCIR, CLEF, Senseval, DUC, ACE).

O TREC e as suas várias provas ("tracks") será muito referido neste livro, enquanto o SUMMAC e o DUC (Rino & Pardo, neste volume), o MUC (Mota *et al.*, neste volume) e o CLEF (Rocha & Santos, neste volume) também serão será feita mais referência:

O NTCIR (NII-NACSIS Test Collection for IR Systems) é um projecto japonês iniciado em 1998 para avaliação conjunta na Ásia (tratando do japonês, inglês, chinês e coreano) e que vai neste momento na sua quarta edição, com cinco competições diferentes (além de RI monolingue e cruzada, sumarização, RI na Web e procura em patentes).

O Senseval, que vai neste momento na sua terceira edição, testa a desambiguação de sentidos de palavras ambíguas, sendo o recurso dourado o conjunto de sentidos num dado dicionário (Kilgarriff & Rosenzweig, 2000).

As Morpholympics (dedicadas à avaliação de analisadores morfológicos para o alemão, Hausser, 1996), e o Parseval (comparando analisadores sintácticos do inglês, Black *et al.*, 2001) apenas se realizaram uma vez. O que não significa que avançassem menos o estado da área ou que a sua realização não constituísse um marco, ou fonte de inspiração, para trabalho futuro. Muitas vezes, a não repetição do acontecimento deve-se simplesmente à falta de uma infra-estrutura permanente. Outras vezes, representa o reconhecimento de que uma repetição não terá vantagens suficientes para o que se quis descobrir, compilar, ou investigar.

3.1 Contrapondo o TREC e o MUC

Estas duas iniciativas, ambas americanas, congregaram comunidades diferentes (PLN por um lado e RI por outro). De facto, é curioso observar que Hirschman (1998a) junta o ATIS[6] e o MUC como avaliações de compreensão de linguagem ("language understanding") e não menciona o TREC nesse contexto. Noutro artigo, Hirschmann (1998c:298f) comenta que uma das principais diferenças entre o TREC e o MUC é a maturidade da tecnologia: em recolha de informação (RI) havia sistemas reais, e daí a

[6] Avaliação conjunta en reconhecimento de fala organizada pelo NIST desde 1987, cf. Pallett (1998).

forte participação da indústria no TREC; enquanto que, em PLN, foi apenas a comunidade académica que se juntou no MUC, ainda longe de produzir sistemas capazes de chegar ao mercado.

Se isto se verificava em 1998, neste momento o TREC está a aumentar as suas ambições, englobando áreas tradicionalmente consideradas PLN como é o caso da resposta automática a perguntas (RAP). Apesar de cada vez mais investigadores (dos dois campos) tentarem aplicar PLN a RI, estou convencida de que permanece ainda alguma fricção entre as duas comunidades e paradigmas.

Deve contudo referir-se que, tendo acabado o MUC, surgiu o ACE, com ambições ainda maiores – a detecção de entidades, relações e acontecimentos em texto, som e imagem (Doddington *et al.*, 2004).

3.2 O modelo do francês

É interessante verificar que a comunidade do processamento do francês (cuja realidade linguística é parecida com a do português, no sentido de haver vários países independentes, espalhados pelo mundo, com a mesma língua oficial) seguiu um modelo diferente do nosso, ainda que com o mesmo objectivo: aplicar ao francês este paradigma já testado e elogiado para o inglês.

Assim, o ciclo Amaryllis (Coret *et al.*, 2000) começou com uma chamada para organizadores em 1994, e apenas dois anos mais tarde uma chamada para participantes. Aqui vemos que à parte "conjunta" não foi dado o relevo que nós demos para o português.

Por outro lado, é inegável a vantagem de ter vários organizadores diferentes para áreas diferentes, como aconteceu na comunidade de língua francesa, que em 1998 congregava na rede FRANCIL 9 países com 69 participantes distintos (Mariani, 1998).

No modelo francês, além disso, a participação nas avaliações conjuntas (ou seja, a adaptação dos sistemas ao modo de avaliação) era financiada.

Em Simões & Almeida (neste volume) será apresentada uma destas avaliações conjuntas, nomeadamente o ARCADE, centrado no alinhamento de textos traduzidos entre francês-inglês (Véronis & Langlais, 2000).

4. A implementação de uma actividade de avaliação conjunta

4.1 Cartografia do problema

Uma questão que me parece pertinente mas que não tem sido, até agora, sublinhada pelas avaliações conjuntas internacionais, é a possibilidade de medir o problema, antes mesmo de avaliar o desempenho da solução.

De facto, da forma como a Linguateca apresentou inicialmente a proposta, havia vantagem em fazer avaliações conjuntas (ou ensaios) com o objectivo de definir o problema e obter uma medição da situação, antes mesmo de implementar os sistemas que o tentassem resolver. Pensamos, aliás, que apelar para as pessoas interessadas, mesmo que ainda não tivessem sistemas "prontos", foi uma das razões que levou a uma resposta tão positiva por parte da comunidade.

Como tenho referido noutros artigos, sem sabermos a dificuldade (ou entropia) de uma dada tarefa, uma avaliação não pode ser útil, e os números são ininterpretáveis. Por exemplo,

- que trabalho dá alinhar textos paralelos? Um procedimento automático melhora ou também pode piorar? (Santos & Oksefjell, 2000)
- se, em certos casos, é possível interpretar um sinal de pontuação como uma conjunção, deve medir-se a competência de um sistema sobre TODOS os sinais de pontuação?
- que hipóteses estão subjacentes à etiquetação morfossintáctica e qual o significado das medidas de precisão? (Santos & Gasperin, 2002; Santos, 2003)

Outra observação extremamente relevante diz respeito à língua: Palmer & Day (1997:193), após terem experimentado um dado algoritmo para o reconhecimento de entidades mencionadas em várias línguas, argumentam que os RMDs para línguas diferentes e para colecções diferentes podem não ser comparáveis. Ou seja, valores iguais podem significar desempenhos diferentes.

4.2 A definição da tarefa

É fundamental, ao organizar uma avaliação conjunta, definir uma tarefa o mais possível neutral e compreensível por pessoas, mesmo que não seja necessariamente de utilidade para o utilizador final. Ou seja, tem de estar clara e bem definida, e poder ser compreendida (e repetida) por um conjunto de pessoas diferentes.

Veja-se por exemplo a atenção dada ao processo de criar tópicos "naturais", mudado do TREC-3 para o TREC-4, é discutido por Harman (1996). De facto, Hirschmann (1998b) até sugere que uma boa alternativa ao MUC seria fazer uma avaliação conjunta de sistemas de compreensão de texto usando os materiais pedagógicos já existentes para ensinar língua estrangeira (ou a própria) a seres humanos.

Por outro lado, a avaliação de uma tarefa bem definida não tem de ser unívoca, no sentido de comparar cegamente com um recurso dourado: Katz & Arosio (2001) demonstram que, muitas vezes, é preciso ir mais além do que uma simples concordância entre diversas formas distintas de anotar: ao marcar relações temporais entre acontecimentos num texto, a consistência semântica é diferente (e mais abrangente) do que identidade, e obviamente muito mais importante.[7]

4.3 Recursos

Na prática, há duas formas de organizar uma avaliação conjunta:

- criar um recurso dourado, feito total ou parcialmente (revendo algum resultado automático) por pessoas a desempenhar a tarefa que se pretende ser feita automaticamente mais tarde, tal como aconteceu no MUC, no Senseval ou nas Morfolimpíadas; ou usando para tal um recurso já criado anteriormente, como aconteceu com o Penn Treebank no Parseval (Black *et al.*, 1993);
- ou confiar de algum modo nos sistemas participantes, cuja soma de resultados, ordenados, redunda num monte ("pool") e, depois, apenas ajuizar os resultados desse monte, como é o caso do TREC ou do CLEF. (Para discussão das formas e consequências deste amontoamento, veja-se Zobel (1998) e Braschler & Peters (2003).)

Relativamente a uma avaliação com base num recurso dourado, Will (1993) relata que, no preenchimento das chaves do MUC 5[8], mesmo depois de várias sessões de treino, houve 33% de erro (ou discordância) quando preenchidas por anotadores humanos (qualificados como analistas, mas não peritos em micro-electrónica). Este

---

[7] Por exemplo, se alguém marcou "A antes de B" e "B antes de C", e outros marcaram "C depois de A" e "B depois de A", nenhuma das alternativas deve ser privilegiada.

[8] Cuja tarefa era a extracção de informação de textos científicos, nomeadamente de microelectrónica.

artigo descreve como comparar o desempenho de vários anotadores, e sobretudo a necessidade de procedimentos complexos para criar chaves o mais possível correctas usando mais do que um profissional. Este é um problema fulcral para todos os casos em que se pretende obter um recurso dourado: Brants (2000) discute medidas de concordância entre anotadores de análise sintáctica; Setzer (2001), no âmbito da anotação semântica, levanta a problemática da dependência entre tarefas de anotação e o espaço das suas medidas.[9] Tinsler & Weiss (2000) discutem os fundamentos matemáticos deste tipo de estudos.

Por outro lado, um esforço considerável tem sido investido para validar a abordagem seguida no TREC (o chamado modelo de Cranfield), visto que, conforme mencionado por Voorhees (1998), desde o início os seus detractores não têm deixado de pôr em causa a noção de relevância. Harman (1996) verificou que, apesar do julgamento de relevância ser incompleto (dado que apenas um subconjunto dos documentos é revisto pelos juízes), aumentar o tamanho do monte julgado redundava em muito poucos novos documentos relevantes, nos primeiros TREC, enquanto que Voorhees (1996), verificou, além disso, que a ordem relativa dos sistemas era estável mesmo com diferenças de opinião no que respeita a relevância.

4.4 Medidas

As medidas por que se pauta uma avaliação conjunta são a sua pedra de toque. Conforme referido acima, medir só faz sentido se a medida for adequada, i.e., se os valores numéricos apresentarem uma relação forte com as qualidades que se pretende aferir.

Mais ainda, e em ambos as estratégias de construção de recursos (recurso dourado ou amontoamento), é preciso garantir que as medidas feitas com base nesses recursos são confiáveis e estatisticamente relevantes. No caso, além disso, de terem sido usadas pessoas para criar os resultados "certos", é preciso ter uma estimativa do erro introduzido pelo factor humano, e da intersubjectividade máxima, ou seja, até que ponto é que há concordância entre várias pessoas a fazer a mesma tarefa.

Buckley & Voorhees (2000) estudaram as várias medidas empregues na avaliação de RI quanto à sua relação com o tamanho da experiência (quantos tópicos

---

[9] Por exemplo, se duas tarefas se fazem em série, o intervalo possível para valores de concordância na segunda tarefa torna-se drasticamente reduzido pelo intervalo da primeira.

usados) e com as diferenças entre a pontuação obtidos pelos sistemas. Em 2000, o sistema de avaliação do TREC produzia 85 números baseados numa vintena de medidas diferentes. A interpretação dos resultados exigia portanto informação sobre qual o erro associado a cada medida.[10]

Em relação à RAP, foi feito um estudo semelhante ao relatado por Voorhees (1998), para indagar se o método de avaliação era justo no TREC Q&A, e para investigar o número mínimo de perguntas necessário (Voorhees & Tice, 2002). Contudo, devido à forma como o TREC Q&A foi desenhado, não se pode usar os resultados obtidos automaticamente pelos sistemas como um recurso de treino para futuras edições, devido ao facto de os excertos de 50 bytes não terem uma semântica, ou seja, são considerados certos se incluírem a resposta, mas a forma de os obter é irrelevante e varia drasticamente de sistema para sistema.

A relação entre as diferenças nas medidas e a importância ou interesse dos casos resolvidos, ou não, pelos sistemas é também pertinente: Vorhees & Tice (2000b) notam que nem todos os casos são iguais, e que não se faz justiça a um sistema se se dá o mesmo peso às respostas fáceis e às difíceis. Assim, os casos de discordância entre os juízes humanos e o programa julgador automático que implementaram acabam por ser, precisamente, aqueles em que a resposta é mais complicada, e portanto quando é mais necessário ter um julgamento humano...

Mas, salientemos aqui, que todos estes problemas, extremamente pertinentes, só podem ser postos e discutidos depois de haver, pelo menos, uma sessão de avaliação conjunta, para que os dados possam ser trabalhados e as condições da avaliação conjunta avaliadas e melhoradas. Antes disso, contudo, é preciso definir que problema(s) atacar, conforme descrito na secção 4.1 acima.

5. O AvalON: um modelo do português

A história do AvalON[11] começou na discussão do futuro da área e na inquirição sobre que medidas podiam ser tomadas para o avanço desta (Santos, 1999a), sendo uma das

---

[10] Um exemplo do resultado deste estudo foi a compreensão de que a medida "precisão nos 30 primeiros documentos" tem um erro associado duas vezes maior do que a precisão média, ou seja, é preciso avaliar o dobro de perguntas para ter a mesma confiança nos resultados produzidos com esta medida.

[11] Nome guarda-chuva que demos às avaliações conjuntas em geral para o português.

propostas a aplicação deste paradigma. A própria designação de "avaliação conjunta" foi proposta em Santos (1999b). Com a entrada em vigor do projecto Centro de Recursos distribuído para o Processamento da Língua Portuguesa (mais tarde baptizado Linguateca) em 2000, realizaram-se as condições para a primeira avaliação conjunta do português.[12]

De facto, uma característica deste modelo que se poderia mencionar logo à cabeça é o enorme esforço organizativo que nem sempre é possível despender, sobretudo não havendo nos nossos países instituições estatais com este papel e com pessoal afecto a essas actividades. A Linguateca tem feito esse papel em alguns casos (apresentados no presente livro), mas possui um pessoal e um horizonte temporal muito reduzidos, que não são comparáveis aos do NIST (agência americana para a definição e manutenção de padrões) ou da DARPA (agência americana de financiamento de investigação ligada à defesa nacional).[13]

Após um período activo de disseminação da ideia, e da tentativa de sensibilização de todos os potenciais interessados, incluindo a organização de um encontro preparatório em 2002[14], as Morfolimpíadas começaram a tomar forma, assim como outras propostas foram chegando à mesa ou sendo ideadas. (Veja-se Santos & Rocha (2003), assim com o sítio da Linguateca para uma descrição desse processo.)

Neste livro, apresentamos o primeiro passo do paradigma da avaliação conjunta para o português. Embora documentemos principalmente as Morfolimpíadas (capítulos 2 a 11), e relatemos também uma avaliação conjunta de RI e outra de RAP para o português no âmbito do CLEF (capítulo 212, mais de trinta outras avaliações conjuntas esperam a altura de serem realizadas, a avaliar pelo interesse inicial.

Para que isto se realize, temos de esperar que este livro permita demonstrar às agências de financiamento brasileiras e portuguesas que actividades desta índole são

---

[12] A Linguateca não tem como único objectivo organizar avaliações conjuntas, mas a avaliação é parte integrante do seu modelo IRA: Informação, Recursos e Avaliação, Santos (2000b, 2002).

[13] Para apresentar dados concretos, Hirschman (1998b) revela que a organização de cada MUC custou duas a quatro pessoas-ano. Paroubek & Blasband (1999) apresentam dados financeiros sobre várias avaliações conjuntas europeias. Kilgarriff (2003), contudo, defende para o Senseval uma forma de organização voluntária não-financiada, parecida com a realizada pela Linguateca.

[14] Com mais ou menos o mesmo âmbito e objectivo que o encontro de 1988 documentado por Palmer & Finin (1990).

fulcrais no desenvolvimento concertado de uma área de investigação e que deviam ser implementados mecanismos para proporcionar a sua organização (ou o apoio a essa organização) numa base permanente, à semelhança do que é feito noutras comunidades.

As próprias Morfolimpíadas apenas terão verdadeiro impacto se, daqui por diante, as pessoas usarem os recursos nelas produzidos para

1) identificarem os problemas
2) medirem os seus sistemas
3) melhorarem o estado global da área

Por outro lado, a maior parte das outras propostas apresentadas neste livro são embrionárias, dado que ainda não reflectem uma real participação da comunidade e, portanto, avanço na área. Que sejam, pelo menos, inspiradoras para os grupos vindouros, que possam partir daqui, quer criticando, quer concordando, com o pouco (mas tanto) que já foi feito.

O que ainda falta completamente, temos de repeti-lo, é o interesse dos sistemas comerciais e das agências de financiamento. Esperemos que este livro contribua para colmatar essa dupla ausência. Mas, se conseguirmos despertar no leitor o respeito e interesse por este paradigma, aliciando-o para participante ou pelo menos observador interessado de futuras avaliações conjuntas, o livro – e este artigo – terá cumprido o seu principal papel.