INFORMÁTICOS, LINGUISTAS E LINGUAGENS

COMPUTER SCIENCE, LINGUISTS AND LANGUAGES

Alberto Simões
CENTRO DE ESTUDOS HUMANÍSTICOS, UNIVERSIDADE DO MINHO
ambs@ilch.uminho.pt

1. Prólogo

Existe uma espécie de *Guerra Santa* há algum tempo entre investigadores da área das ciências da computação (a que vou chamar abusivamente de informáticos) e investigadores da área das ciências da língua (a que vou chamar abusivamente de linguistas) porque os primeiros se têm aventurado em tarefas que habitualmente eram realizadas pelos segundos. Estas incursões levam a que tarefas que habitualmente demoram meses a realizar de forma manual sejam automatizadas e realizadas rapidamente, com a ajuda de um programa computacional. Tipicamente, quando estes trabalhos são apresentados em conferências habitualmente frequentadas por linguistas, são alvo de grandes críticas pela *falta de correção* do resultado obtido. O que pretendo apresentar neste documento são as razões que me parecem levar a este comportamento, e discutir o que é possível alcançar se informáticos e linguistas conseguirem perceber os pontos de vista e objectivos de cada um deles.

O mais natural é que, ao ser interrogado sobre o suposto trabalho de linguista, um informático comum imagine alguém de óculos grossos debruçado sobre um texto (muito frequentemente literário), a sublinhar palavras e a tirar notas. Do mesmo modo, muitos linguistas irão descrever um informático como um ser antissocial que é capaz de comunicar por uma linguagem estranha com computadores. Estes são estereótipos, mas que por vezes não fogem muito da verdade. Tudo depende do informático ou do linguista que se escolha para analisar.

Para esta deambulação, que não passa de uma visão muito pessoal, como informático, da forma como me senti nos primeiros tempos em que trabalhei em processamento de linguagem natural, irei seguir a seguinte abordagem: Em primeiro lugar discutirei um pouco sobre a questão das linguagens, naturais e artificiais. Depois será apresentado o conceito de completude subjacente à análise da língua, quando esta é realizada por informáticos ou por linguistas. Segue-se uma análise à questão da terminologia e da sua relevância para a comunicação entre informáticos e linguistas, que culmina com a apresentação de tarefas em que linguistas podem tirar partido dos informáticos, e vice-versa.

2. Linguagens

Os informáticos podem ser divididos em diferentes tipos, de acordo com a sua atividade principal, desde os administradores de sistemas, que habitualmente dão a fama aos restantes, de serem resmungões e preguiçosos, até aos engenheiros de software que raramente tocam nos computadores, e passam a vida a desenhar diagramas. Aqueles que habitualmente lidam com o processamento da língua, sendo eles investigadores ou programadores de profissão, passam grande parte do seu tempo a escrever e a analisar programas informáticos.

É aqui que não posso deixar de fazer uma ponte entre este tipo de informáticos e os linguistas: as linguagens. As linguagens são habitualmente divididas em dois grandes grupos: as linguagens ditas naturais, que nasceram e cresceram à medida que o ser humano as foi utilizando para comunicar; e as linguagens ditas artificiais, que foram criadas, de forma controlada, pelo ser humano (Okrent, 2010) (Adams, 2011).

Exemplos das linguagens naturais são praticamente todas as línguas que o ser humano usa para comunicar, daí que muitas vezes em vez de linguagens usemos o termo "língua" para nos referirmos a elas. Por outro lado, uma língua como o Esperanto, é uma linguagem artificial, e é a única língua artificial conhecida realmente usada para a comunicação entre seres humanos. Outras línguas artificiais foram desenvolvidas com outros propósitos. Por exemplo, na tradução automática foi definida uma representação dita semântica das frases que foi batizada de "*interlíngua*" (Dorr, 1992). No entanto, esta linguagem artificial foi desenvolvida para ser usada por máquinas.

Outro grande grupo de linguagens artificiais são usadas pelos informáticos, e são habitualmente designadas de linguagens de programação, ou linguagens formais. Exemplos destas linguagens não são só as ditas de programação (Ber-

gin & Gibson, 1996), como o C, Assembler, Java, Perl, Python ou Ruby, mas também as ditas de anotação, como o LaTeX, o HTML ou o CSS. Estas são linguagens com um léxico muito reduzido e com uma semântica associada não ambígua, o que permite que as máquinas, determinísticas, sejam capazes de as interpretar e executar.

3. Linguagens e Arte

Será natural que se associe a arte às linguagens naturais. Os romances, escritos em prosa, são, em muitos casos, exemplos de arte. Do outro lado, a poesia será por muitos considerada ainda mais artística. A pergunta que se pode fazer é: *será possível usar uma linguagem formal e criar arte?*

Suponho que tudo dependerá do ponto de vista de cada um sobre o que devemos considerar arte. No entanto, existem exemplos que facilmente podem ser considerados como tal. E, na verdade, parece-me que o processo de usar uma linguagem formal para este tipo de tarefa não será menos complicado que o uso de uma linguagem natural.

Considere-se o seguinte exemplo de poesia "tradicional", por *W.B. Yeats* de 1916:

> *Though leaves are many, the root is one;*
> *Through all the lying days of my youth*
> *I swayed my leaves and flowers in the sun;*
> *Now I may wither into the truth.*
> (Yeats, *Responsibilities and Other Poems*, 1916: 103[1])

Num concurso de poesia usando a linguagem de programação Perl, *Wayne Myers* traduziu este poema como:

```
while ($leaves > 1) {
    $root = 1;
}
foreach($lyingdays{'myyouth'}) {
    sway($leaves, $flowers);
}
```

1 Yeats, W.B. (1916), *Responsibilites and Other Poems*, New York: Macmillan.

```
while ($i > $truth) {
    $i--;
}
sub sway {
    my ($leaves, $flowers) = @_;
    die unless $^O =~ /sun/i;
}
```

Embora a sintaxe e semântica associada a esta linguagem de programação não seja muito clara para alguém que não a conheça, é relativamente fácil de perceber pelo menos o primeiro bloco, que se poderá ler como: *enquanto as folhas (leaves) forem mais que uma, então a raíz (root) é uma*.

Este tipo de exercício pode ser visto como a extração de uma semântica do poema original, o que não é propriamente o tipo de arte mais habitual.

Um outro exemplo, desta vez escrito diretamente na linguagem de programação Perl, por *perlaintdead*:

```
#!/usr/bin/perl
study $science;
open $doors,">","success.txt";
BEGIN aNewLife() and use Hard::Work;
push for(;;){},"a brighter future";
do not do $meth;
do not return 2, -e$x-$girlfriend;
shift @yourPerspective;
exit $fear;
```

Neste excerto, a primeira linha, típica dos programas em Perl, não faz parte do poema. De seguida o texto lê-se de forma mais ou menos simples: *"estudar ciência, abrir portas para o sucesso; iniciar uma nova vida cheia de trabalho duro"*. Este excerto é interessante por usar uma sintaxe válida da linguagem Perl, mas infelizmente a semântica não está correta, e o programa não executa.

Outros exemplos, para além de poderem ser lidos como os anteriores, no código em que foram escritos, também executam. Segue-se um exemplo apresentado por *hdb* no sítio *PerlMonks*[2]:

---

2 http://www.perlmonks.org/?node_id=1028045

```
use strict; # just kidding
my%hat;my@cat=split/ /,<DATA>;
while(@cat){
  my($cat=>$hat)=splice@cat,0,2;
  $hat{$cat}=$hat>$hat{$cat}?$hat:$hat{$cat};
}
print join" ",%hat; # and that was that
__DATA__
a.txt 4 b.txt 3 z.txt 1 a.txt 5 b.txt 2 b.txt 4 z.txt 2
```

Neste caso a semântica do programa, como texto da língua inglesa, é praticamente nula. No entanto o jogo de palavras misturado tem um efeito interessante e, para além disso, o programa é correto e é capaz de produzir um resultado.

Se os informáticos são capazes deste tipo de arte não poderão estar tão longe dos colegas das artes ou humanidades.

4. Os 98%

Falaremos agora de um assunto bem mais importante e sério: quais os objetivos de informáticos e de linguistas ao analisar a língua, e porque é que habitualmente não se entendem.

Ao analisar a língua, alguns linguistas gostam de encontrar e classificar fenómenos linguísticos, sejam ocorrências de terminado verbo, de determinada construção, seja para inferir o sentido das palavras de acordo com o seu contexto. Por vezes, o que realmente é interessante para o linguista não é encontrar aquelas formas ou aqueles sentidos que vêm registados nas gramáticas ou dicionários. O seu interesse é encontrar as situações fora do comum. Ora, no mundo real, essas situações correspondem a cerca de 2% das ocorrências analisadas. Por outro lado, outros trabalham na generalização ou análise da língua como um todo, e estes embora não se centrem em apenas 2% dos fenómenos linguísticos, não se contentam com menos que 100% da língua.

Por outro lado, os informáticos gostam de automatizar. Na definição de informática referem-se "*processos racionais e automáticos para o processamento de dados*". Para que um processo seja automatizável é necessário generalizá-lo. Este processo de generalização leva a que as situações menos comuns ou normais não sejam encontradas, e portanto, não sejam processadas. Ou seja, os

informáticos não se importam se conseguirem sistematizar 98% das situações e ignorar os 2% de situações que não são capazes de generalizar.

Esta visão pragmática ou funcional do processamento da língua é imprescindível para o tratamento automático da língua mas aborrece muitos linguistas que argumentam não ser relevante um processo que não seja capaz de lidar com todas as possibilidades da língua.

O que é importante realçar a este respeito é que ambas as visões deste problema são relevantes, mas que têm objetivos diferentes. A automatização de um processo visa o processamento de grandes quantidades de documentos, processo esse que não é possível de ser realizado de forma manual. Talvez se conseguisse automatizar mais 0.5% das situações. O problema é que o tempo investido para esse aumento na cobertura pode ser demasiado dispendioso (seja a nível monetário, temporal ou de poder computacional). Do outro lado, os 2% que os linguistas querem analisar, são situações esporádicas ou científicas, onde o custo monetário ou temporal não é relevante.

Na secção 6 apresento uma visão sobre a cooperação entre informáticos e linguistas que assenta, exatamente, nesta questão dos 2% vs 98% de cobertura.

5. Terminologia

Um dos primeiros obstáculos na cooperação entre informáticos e linguistas é a terminologia. Sem dúvida que a terminologia é importante, ou não se faziam terminologias multilingues, nem se insistia tanto com os alunos de tradução que um dos principais cuidados na tradução técnica ou científica é a tradução correta da terminologia. Também na escrita de artigos científicos a terminologia da área em causa deve ser respeitada.

Sem dúvida que noutras situações a terminologia não é tão relevante, ou então, o seu uso tem de ser comedido. Considere-se uma notícia sobre a descoberta de um novo medicamente. Se a notícia for apresentada usando a terminologia específica da área poucos serão aqueles que a conseguirão compreender.

Se pensarmos na comunicação, como uma necessidade, vemos que a terminologia é prescindível. Ninguém vai à mercearia pedir *"cloreto de sódio"*, ou a um restaurante pedir um bife de *"Bos taurus"*. Noutra situação, quando visitamos um país estrangeiro e não conhecemos a língua, nem sequer precisamos de usar a linguagem oral ou escrita, e com um diagrama ou uma imagem conseguimos fazer o nosso interlocutor perceber o que pretendemos.

Informáticos e linguistas devem ter isto em consideração. A formação de base de ambos é a mesma (ensino nacional obrigatório) mas a formação específica é diferente. Ambos irão conseguir aprender a terminologia das áreas dos respectivos colegas, mas não instantaneamente.

É necessária uma aproximação, em que cada um seja capaz de se exprimir sem usar a terminologia específica de uma área. Para isso, muitas vezes, será preciso explicar o conceito. É possível que esse conceito tenha de ser explicado várias vezes ao longo da cooperação. Mas é assim a aprendizagem.

6. Cooperação

Existem situações em que os informáticos têm interesse na ajuda dos linguistas para tarefas que, para estes últimos, não são muito úteis. Outras situações em que os linguistas podem tirar partido da ajuda dos informáticos, sem que estes últimos possam tirar qualquer proveito. Por fim, existem ainda outras situações, cada vez mais frequentes, em que a cooperação é útil para todos os intervenientes. Neste capítulo apresentarei alguns exemplos referentes a estas três situações.

6.1 Colaboração em projetos informáticos

Um exemplo de projeto informático cuja realização poderia estar comprometida sem a ajuda de linguistas é o caso da indexação de um acervo documental. Na era em que estamos, todas as organizações, sejam elas governamentais, científicas, industriais ou mesmo relativas à saúde, produzem grandes quantidades de documentação. Esta documentação, atualmente, está a ser vista como uma mais-valia, um recurso valioso, que deve ser estudado e preservado. Ora, como sabemos, o papel não é dos suportes com mais longevidade, nem um suporte fácil de manusear. Daí que estas organizações estão a proceder à digitalização e indexação destes documentos. Alguns destes acervos vão sendo tornados públicos, como é o caso do Projeto Gutenberg (Hart, 1971), do Dicionário Aberto (Simões & Farinha, 2011), do Mutopia (2013), ou outros. Outros projetos, por serem uma mais-valia para a organização que os produziu, não são disponibilizados gratuitamente, mas são preparados e armazenados.

Neste processo o conhecimento linguístico é imprescindível. Consideremos a pesquisa por palavras-chave nestes documentos. Sem a existência de

uma ferramenta capaz de lematizar as palavras digitadas pelo utilizador, bem como as palavras dos documentos, a pesquisa nestes documentos era muito pouco eficaz.

6.2 Colaboração em projetos linguísticos

Há um conjunto de trabalhos puramente linguísticos que se tornavam demorados e cansativos sem o apoio de informáticos, como sejam a análise de construções frásicas concretas, ou do contexto em que determinadas palavras são usadas. Por exemplo, a compilação de um dicionário (Sinclair, 1988) (Gómez Guinovart et al., 2012) com recurso a corpora, ou o estudo de determinadas construções sintáticas (Araújo, 2008), são trabalhos cuja realização manual é tarefa muito árdua. Com o desenvolvimento de aplicações que permitem a consulta de concordâncias em corpora (por exemplo, o motor de indexação de corpora *Corpus Workbench* (Evert & Hardie, 2011), ou as interfaces de pesquisa em corpora AC/DC (Santos, 2014), OPUS (Tiedemann, 2012) e Per-Fide (Araújo et al., 2010)[3]) este processo torna-se simples e muito mais eficaz, ajudando a encontrar de forma mais rápida os ditos 2% que são interessantes para o problema em análise.

6.3 Colaboração em projetos interdisciplinares

Embora os exemplos anteriores sejam válidos e interessantes, habitualmente para aqueles que recebem a ajuda, sejam eles linguistas ou informáticos, existem muitos outros projetos, muito mais interessantes, que podem lucrar com o trabalho de ambos. Aqui apresento apenas três pequenos exemplos de investigação em curso com recurso a equipas interdisciplinares, não só de informáticos e linguistas, mas também de outros especialistas, como especialistas em estatística.

- Tradução automática
  A história relativa à evolução e investigação em tradução automática é o claro exemplo da necessidade de equipas multidisciplinares. Tudo surgiu com informáticos que tinham algum interesse na língua, e deci-

3 Verdade seja dita que em muitos destes trabalhos os informáticos acabam por tirar partido das ferramentas que desenvolvem para outras finalidades que lhes sejam úteis.

diram colocar em prática algumas ideias já publicadas por linguistas, nomeadamente o conceito de interlíngua, ou seja, a possibilidade de representar qualquer frase de qualquer língua numa linguagem artificial. As primeiras tentativas na implementação de sistemas baseados neste conceito mostraram que havia uma área de investigação relevante. Estes sistemas evoluíram com algumas técnicas de engenharia e muitas poucas alterações do ponto de vista linguístico.

Com o advento dos sistemas computacionais com grandes capacidades de armazenamento e computação assistiu-se ao abandono dos métodos ditos simbólicos, que lidavam com estruturas e regras da língua, para métodos puramente estatísticos (Hutchins & Somers, 1992) (Brown et al., 1990). Nesta altura, as equipas multidisciplinares perderam os linguistas, que foram substituídos por matemáticos e estatísticos. É um facto que foi nesta altura que surgiram alguns dos grandes sistemas de tradução, como é o caso do bem conhecido sistema online da Google.

Os resultados obtidos não melhoraram apenas pelo tipo de abordagem em causa mas pela quantidade de textos disponíveis que puderam ser analisados e que permitiram aos sistemas aumentar o seu léxico e base de estruturas gramaticais, de um modo que não poderia ser obtido usando informação introduzida de forma manual.

A demonstração de que os linguistas são parte imprescindível nesta tarefa é ter-se regressado à incorporação de regras simbólicas para melhorar e corrigir os resultados obtidos pela tradução puramente estatística (Riezler & Maxwell III, 2006).

- Sumarização

  A sumarização tem como objectivo obter informação de um texto, selecionar aquela que é relevante, e construir um pequeno texto que resuma o texto original. Atualmente muitos sistemas baseiam-se apenas em informação estatística para selecionar quais as frases mais relevantes que serão apresentadas sem qualquer alteração no resumo proposto. Este tipo de funcionamento é claramente ineficaz já que assume que o texto original inclui frases que não acrescentam informação e podem ser simplesmente ignoradas.

  Nesta situação o trabalho dos linguistas é bastante importante para ajudar os informáticos a delinear algoritmos ou pelo menos heurísticas para detectar e compreender quais as partes relevantes de um texto e,

mais tarde, ajudar a recuperar essa informação num texto legível e correto (Barzilay & Elhadad, 1997) (Carbonell & Goldstein, 1998).

- Resposta a perguntas
  Esta área tem bastantes semelhanças com a anterior. O objectivo é o desenvolvimento de uma aplicação capaz de entender uma pergunta introduzida pelo utilizador em linguagem natural, saber como procurar uma resposta num conjunto de documentos ou bases de dados disponíveis e, com base nos factos recolhidos produzir uma resposta também em linguagem natural. Este exercício é muitas vezes estendido, considerando que a pergunta e resposta podem estar escrita numa língua diferente da usada no documento que descreve o facto que se procura (Lin & Pantel, 2001) (Clarke et al. 2001) (Radev et al., 2002).

- Análise do discurso
  A análise de discurso ou minagem de opinião é uma área relativamente recente que procura tirar partido da quantidade de notícias e a velocidade a que elas são comentadas na rede para interpretar diferentes aspectos como sejam a popularidade de um cantor ou de uma marca, a decisão de determinado político ou a evolução económica de um país ou de uma empresa. Mais uma vez se não se recorrer a conhecimento linguístico será muito fácil cometer erros grosseiros, desde a má interpretação de adjetivos positivos usados em contextos depreciativos, como em "a empresa x está bem mal" até situações de pura ironia (Drury *et al.*, 2012) (Liu, 2010) (Agarwal et al., 2011).

7. Epílogo

A constituição de equipas multidisciplinares é deveras importante. Para isso é preciso sabermos comunicar e compreender diferentes visões do mundo. No caso concreto dos linguistas, um dos parceiros mais naturais serão, sem dúvida, os informáticos. No entanto não nos podemos limitar a estes. Outras áreas, como a psicologia, educação ou mesmo a medicina, definirão outras pontes igualmente interessantes e relevantes. Do mesmo modo, os informáticos podem lucrar muito com a ajuda dos linguistas, nomeadamente aqueles cujo trabalho é centrado no processamento de informação descrita em linguagem

natural. Mas outras pontes também são possíveis, como a electrónica, a física ou a química.

A colaboração interdisciplinar entre linguistas e informáticos tem surgido num conjunto de diferentes áreas de conhecimento, como sejam o processamento de linguagem natural, a linguística de corpora ou computacional, lexicografia computacional, entre outros. E ainda muito há para fazermos em conjunto no futuro.

Referências

Adams, Michael (2011), *From Elvish to Klingon: Exporing Invented Languages*, Oxford: OUP.

Agarwal, Apoorv, Boyi Xie, Ilia Vovsha, Owen Rambow & Rebecca Passonneau (2011), *Sentiment Analysis of Twitter Data*, *in* Proceedings of the Workshop on Languages in Social Media (LSM '11), Stroudsburg, PA, USA: Association for Computational Linguistics, pp: 30-38.

Araújo, Sílvia (2008), *Entre l'actif et le passif: se faire/fazer-se: syntaxe, sémantique et pragmatique comparées français-portugais*, Tese de Doutoramento, Universidade do Minho.

Araújo, Sílvia, José João Almeida, Alberto Simões, & Idalete Dias (2010), *Apresentação do projecto Per-Fide: Paralelizando o português com seis outras línguas*. Linguamática, 2(2):71-74, Junho.

Barzilay, Regina & Michael Elhadad (1997), *Using lexical chains for text summarization*, *in* Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, pp. 10-17.

Bergin, Thomas J. & Richard G. Gibson (1996), *History of Programming Languages*, ACM Press, Addison Wesley.

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vicent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer & Paul S. Roossin, *A statistical approach to machine translation*, Computational Linguistics, 16(2): 79-85, 1990.

Carbonell, Jaime & Jade Goldstein (1998), *The use of MMR, diversity-based reranking for reordering documents and producing sumaries*, *in* proceedings of Special Interest Group on Information Retrieval, pp. 335-336.

Clarke, Charles L. A. Clarke, Gordon V. Cormack & Thomas R. Lynam (2001), *Exploiting Redundancy in Question Answering*, *in* Proceedings of Special Interest Group on Information Retrieval, ACM Press, pp. 358-365.

Dorr, Bonnie (1992), *Parameterization of the interlingua in machine translation*. In Proceedings of the 14th conference on Computational linguistics - Volume 2 (COLING '92), Vol. 2., Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 624-630.

Drury, Brett, Luís Torgo & José João Almeida (2012), *Classifying News Stories with a Constrained Learning Strategy to Estimate the Direction of a Market Index*, International Journal of Computer Science and Application. 9 (1): pp.1-22.

Evert, Stefan & Andrew Hardie (2011), *Twentry-first century Corpus Workbench: Updating a query architecture of the new millennium*, *in* Proceedings of the Corpus Linguistics 2011 Conference, University of Birmingham, UK.

Gómez Guinovart, Xavier, Alberto Álvarez Lugrís & Eva Díaz Rodríguez (2012), *Dicionario moderno inglés-galego*. 2.0 Editora: Ames.

Hart, Michael (1971), *Project Gutenberg*, Project Gutenberg.

Hutching, John & Harold L. Somers (1992), *An Introduction do Machine Translation*, London: Academic Press.

Lin, Dekang & Patrick Pantel (2001), *Discovery of Inference Rules for Question Answering*, Natural Language Engineering,7: 343--360.

Liu, Bing (2010), *Sentiment analysis and subjectivity*, Handbook of Natural Language Processing, Second Edition, Abingdon: Taylor and Francis Group.

Mutopia Project (2013), *The Mutopia Project: Free sheet music for everyone*, disponível em http://www.mutopiaproject.org/ [consultado em setembro 2013].

Okrent, Arika (2010), *In the Land of Invented Languages*, New York: Random House Inc.

Radev, Dragomir, Weiguo Fan, Hong Qi, Harris Wu & Amardeep Grewal (2002), *Probabilistic question answering on the Web*, Journal of the American Society for Information Science and Technology, pp. 408-419.

Riezler, Stefan & John T. Maxwell III (2006), *Grammatical Machine Translation*, *in* Proceedings of North American Chapter of the Association for Computational Linguistics, Human Language Technologies, pp. 248-255.

Santos, Diana (2014), "Corpora at Linguateca", *in* Tony Berber Sardinha & Telma São Bento Ferreira (eds.), *Working with Portuguese Corpora*, London, New York and Sydney: Bloomsbury, pp: 219-236.

Simões, Alberto & Rita Farinha (2011), *Dicionário Aberto: um recurso para processamento de linguagem natural*, Vice-Versa, 16:159-171, December.

Sinclair, John (1988), *Colling Cobuild English Language Dictionary*, Colling CoBUILD.

Tiedemann, Jörg (2012), *Parallel Data, Tools and Interfaces in Opus*, *in* Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012).