A utilização de informação linguística em abordagens baseadas em aprendizagem automática para o processo de tradução

Paulo Miguel Torres Duarte QUARESMA[1]

RESUMO

O processo de tradução automática de textos tem sido objecto de abordagens bastante distintas: desde propostas que visam a análise profunda dos textos e a representação da informação num formato independente da Língua, até abordagens baseadas na análise superficial aos textos e às palavras que os compõem, passando por abordagens estatísticas do processo de tradução.
Neste artigo analisa-se uma abordagem mista e integradora ao processo de tradução automática: por um lado, efectua-se uma análise linguística aos textos, efectuando a sua etiquetagem morfo-sintáctica, a identificação de entidades mencionadas, uma análise sintáctica parcial e uma análise semântica parcial; por outro lado, recorre-se a técnicas de aprendizagem automática, baseada em métodos estatísticos, para criar modelos de tradução entre pares de Línguas, utilizando corpora paralelos disponíveis nestas Línguas. A integração entre estas duas metodologias é efectuada através da utilização dos resultados da análise linguística como *input* de métodos de aprendizagem supervisionada.
Esta estratégia de integração de informação linguística com técnicas de aprendizagem automática revelou bons resultados quando aplicada ao problema de classificação automática de textos em Língua Portuguesa e pretende-se, neste artigo, analisar a sua possível aplicação à temática da tradução automática de textos em Língua Portuguesa.

PALAVRAS-CHAVE: Tradução Automática; Aprendizagem automática; Linguística computacional

INTRODUÇÃO

A tradução automática de textos é uma área de investigação que tem sido objecto de análise ao longo das últimas décadas. As abordagens propostas têm sido bastante distintas e têm sofrido variações significativas ao longo do tempo, acompanhando a evolução e os desenvolvimentos das áreas de Linguística Computacional e de Inteligência Artificial: desde propostas que visam a análise profunda dos textos e a representação da informação que veículam num formato independente da Língua,

1 Universidade de Évora, Escola de Ciências e Tecnologia, Departamento de Informática, Rua Romão Ramalho nº 59, 7000 Évora, Portugal, pq@uevora.pt

até abordagens baseadas numa análise superficial aos textos e às palavras que os compõem, passando por abordagens puramente estatísticas do processo de tradução.

Neste trabalho pretende-se analisar e discutir a utilização de uma abordagem mista e integradora ao processo de tradução automático: por um lado, efectua-se uma análise linguística aos textos, efectuando a sua etiquetagem morfo-sintáctica (POS-tagging), a identificação de entidades mencionadas (nomes, instituições, locais, tempo), uma análise sintáctica parcial (identificando sintagmas nominais e verbais e os seus principais constituintes) e uma análise semântica parcial (recorrendo a ontologias externas e à teoria de representação do discurso DRT); por outro lado, recorre-se a técnicas de aprendizagem automática, baseadas em métodos estatísticos, para criar modelos de tradução entre pares de Línguas, utilizando corpora disponíveis nestas Línguas. A integração entre estas duas metodologias é efectuada através da utilização dos resultados da análise linguística como *input* do processo de aprendizagem automática. Esta estratégia de integração revelou bons resultados quando aplicada ao problema de classificação automática de textos em Língua Portuguesa (Gonçalves 2009) e é passível de ser aplicada à temática da tradução automática em geral e de textos em Língua Portuguesa, em particular.

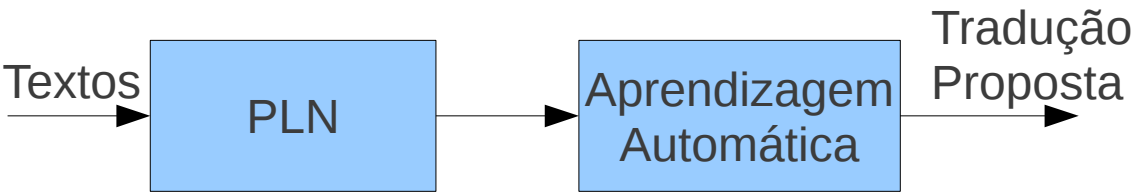
O artigo encontra-se estruturado da seguinte forma: na próxima secção é descrita a arquitectura proposta; na secção seguinte são apresentadas as ferramentas de extracção de informação linguística utilizadas no âmbito do processamento de textos em Língua Portuguesa; posteriormente, é feita uma descrição breve das várias técnicas de aprendizagem automática que podem ser utilizadas no processo de tradução; finalmente, são discutidos alguns dos problemas existentes e possíveis linhas de trabalho e investigação futura.

ARQUITECTURA

De forma a abordar de uma forma integrada o problema da tradução automática de textos, propõe-se o recurso a uma arquitectura modular, em que numa primeira fase os textos são analisados através de ferramentas de processamento de Língua Natural, sendo o resultado dessa análise utilizado por algoritmos de aprendizagem

automática supervisionada, com o objectivo de melhorar os modelos construídos e os resultados obtidos.

De uma forma esquemática, a abordagem proposta pode ser descrita pelo esquema da figura 1:



Os textos, escritos na Língua "origem", são analisados por ferramentas de processamento de Língua Natural, de forma a identificar e a extrair informação linguística (lexical, sintáctica e semântica). Esta informação é utilizada como *input* de um modelo de tradução automático, construído previamente por algoritmos de aprendizagem supervisionada, obtendo-se como *output* uma proposta de tradução na Língua "destino".

Tal como referido, antes de se poder efectuar o processo de tradução automático é necessário realizar o processo de aprendizagem supervisionada, de forma a ser obtido o modelo de tradução a aplicar na segunda fase do processamento referido na figura 1.

Para o processo de aprendizagem supervisionada é utilizada a seguinte metodologia:

a) são aplicadas as ferramentas de processamento de Língua Natural referidas anteriormente (que serão descritas na próxima secção) sobre um conjunto de textos existentes nas Línguas "origem" e "destino";
b) a informação identificada e extraída é representada através de estruturas de dados típicas – grafos, árvores, listas – e é dado como *input* aos algoritmos de aprendizagem automática supervisionada, em conjunto com os pares de tradução Língua "origem" – Língua "destino" existentes no corpus de treino (ver secção "Aprendizagem Automática");
c) o modelo obtido na alínea anterior é salvaguardado e será aplicado para a tradução de novos textos.

É de realçar que esta metodologia implica a necessidade de se ter um corpus paralelo nas duas Línguas, de dimensão razoável e que seja efectuado o seu alinhamento (de uma forma manual ou automática). Requer, ainda, que tenham sido tomadas decisões sobre a granularidade do alinhamento: seja a nível de frases, a nível de segmentos, a nível de entidades ou a nível de palavras.

FERRAMENTAS PARA O PROCESSAMENTO DA LÍNGUA PORTUGUESA

Embora a arquitectura apresentada na secção anterior seja genérica e independente da Língua, e como o foco deste trabalho é a tradução de/para a Língua Portuguesa, nesta secção são descritas algumas ferramentas computacionais existentes para o processamento da Língua Portuguesa, que poderão ser utilizadas para analisar os textos e para produzir como *output* informação linguística relevante para os algoritmos de aprendizagem automática.

1. Etiquetadores morfo-sintácticos – *PoS taggers*

Existem vários etiquetadores morfo-sintácticos para a Língua Portuguesa disponíveis na *web*, com licenças de utilização típicas de software livre (GPL ou LGPL). Uma referência é o software "TreeTagger"[2], desenvolvido por Helmut Schmid (1994), que já foi aplicado a mais de 14 Línguas distintas, incluindo o Português. Uma referência alternativa, mais específica para a Língua Portuguesa, Galego e Castelhano, é o pacote de software FreeLing[3], que inclui várias ferramentas para o processamento de textos em Língua Natural, entre os quais etiquetadores morfo-sintácticos. É de realçar que este pacote de software também efectua um conjunto de operações prévias necessárias à esta tarefa: separação de frases, identificação de termos e análise morfológica, expansão de contracções. Para a tarefa de análise morfológica, o software Jspell[4] (Simões e Almeida, 2001) é também uma referência incontornável.

2 http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
3 http://nlp.lsi.upc.edu/freeling/
4 http://natura.di.uminho.pt/wiki/doku.php?id=ferramentas:jspell

Exemplo: A frase “A Maria leu o livro.” é analisada da seguinte forma pelo analisador FreeLing:

A o DA0FS0 0.667849
Maria maria NP00000 1
leu ler VMIS3S0 0.875
o o DA0MS0 0.944727
livro livro NCMS000 0.977273
. . Fp 1

2. Reconhecimento de Entidades Nomeadas – *NER Named Entity Recognition*

O reconhecimento de entidades nomeadas (NER) é uma tarefa bastante relevante para o processo de tradução automática. Efectivamente, a identificação de, por exemplo, pessoas, entidades, locais e datas permitirá criar agrupamentos coerentes de termos e efectuar a sua análise de uma forma conjunta. O já referido pacote de software FreeLing incorpora módulos que permitem realizar esta tarefa com uma qualidade bastante razoável para a Língua Portuguesa. Um outro sistema disponível também para utilização é o Rembrandt[5] (Cardoso, 2008), que permite, ainda, a identificação de algumas relações entre as entidades identificadas.

Exemplo: O software FreeLing obtém o seguinte resultado para a frase “A Maria leu o livro em Lisboa.”:

A o DA0FS0 0.667849
Maria maria NP00SP0 1
leu ler VMIS3S0 0.875
o o DA0MS0 0.944727
livro livro NCMS000 0.977273
em em SPS00 1
Lisboa lisboa NP00G00 1
. . Fp 1

Realce-se a identificação da etiqueta “SP” (pessoa) no termo “Maria” e da etiqueta “G” (local) no termo “Lisboa.

5 http://xldb.di.fc.ul.pt/Rembrandt/

3. Analisadores sintácticos

A identificação e utilização de segmentos de frases no processo de tradução automática é uma das abordagens que apresenta resultados mais promissores actualmente. Os referidos segmentos poderão ter uma relação directa com o resultado da análise sintáctica das frases, sendo, por exemplo, os sintagmas nominais, preposicionais e verbais dessas frases. Para tal, é necessário ter a capacidade de efectuar uma análise sintáctica dos textos. Para a Língua Portuguesa, existem vários analisadores disponíveis: FreeLing (com um analisador de dependências e um analisador "superficial"[6]); DepPattern[7] (com um analisador de dependências). O analisador PALAVRAS[8] (Bick, 2000) é uma referência reconhecida na área da análise sintáctica da Língua Portuguesa; no entanto, está somente disponível para utilização via *web*, o que inviabiliza a disseminação da sua utilização.

Exemplos:

Os analisadores PALAVRAS e FreeLing obtêm os seguintes resultados para a frase "A Maria leu o livro":

```
PALAVRAS:
SOURCE: Running text
  1. a Maria leu o livro.
  A1
  UTT:cl(fcl)
  .
  |-S:g(np)
  | |-D:pron(det "o" <artd> DET F S)        a
  | |-H:prop("Maria" F S) Maria
  |-P:v(fin "ler" <fmc> PS 3S IND VFIN)     leu
  |-Od:g(np)
    |-D:pron(det "o" <artd> DET M S)        o
    |-H:n("livro" M S)    livro
```

FreeLing:

```
+sn_[
 espec-fs_[
  +j-fs_[
   +(A o DA0FS0 -)
  ]
 ]
 +grup-nom-fs_[
```

6 shallow parser
7 http://gramatica.usc.es/pln/tools/deppattern.html
8 http://beta.visl.sdu.dk/

```
    +w-fs_[
      +(Maria maria NP00000 -)
    ]
  ]
  grup-verb_[
    +verb_[
      +(leu ler VMIS3S0 -)
    ]
  ]
  sn_[
    espec-ms_[
      +j-ms_[
        +(o o DA0MS0 -)
      ]
    ]
    +grup-nom-ms_[
      +n-ms_[
        +(livro livro NCMS000 -)
      ]
    ]
  ]
  F-term_[
    +(. . Fp -)
  ]
]
```

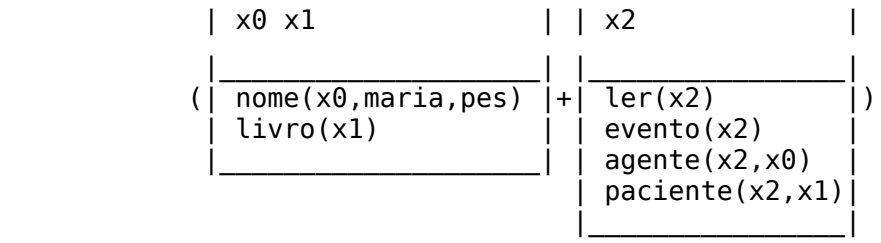
4. Análise semântica

Embora não seja viável efectuar uma análise semântica completa aos textos em Língua Portuguesa, é possível efectuar análises parciais, obtendo anotações semânticas ou, mesmo, estruturas de representação do discurso (DRS). Para o processo de inferência de anotações semânticas pode ser utilizado o software FreeLing; para as estruturas de representação do discurso, pode ser utilizada a ferramenta BOXER[9] (Bos, 2008), que transforma o resultado de análises sintácticas em DRS.

Exemplo:

A frase “A Maria leu o livro” pode ser transformada na seguinte DRS pela ferramenta BOXER (em formato gráfico e com algumas adaptações para a Língua Portugesa, para uma mais fácil visualização):

9 http://svn.ask.it.usyd.edu.au/trac/candc/wiki/boxer



5. Alinhadores de texto

6. Conforme será referido na próxima secção, existem várias metodologias para a construção de sistemas de tradução automática baseados em métodos estatísticos, sendo que esses métodos pressupõem ou incorporam técnicas para o alinhamento automático de textos paralelos escritos em duas Línguas distintas. No entanto, existem alguns alinhadores de texto disponibilizados pela comunidade de investigadores de Língua Portuguesa que podem ser utilizados de uma forma autónoma: CEPRIL[10] e NATools[11].

TRADUÇÃO AUTOMÁTICA ESTATÍSTICA[12]

A tradução automática estatística utiliza técnicas de aprendizagem automática para abordar o problema da tradução de textos. Lopez (2008) efectua uma análise detalhada do estado da arte neste domínio e das várias técnicas utilizadas actualmente.

Uma análise cuidada das várias técnicas utilizadas leva-nos a concluir que o recurso a informação linguística tem vindo a ganhar uma importância crescente nos sistemas de tradução automática estatística, com o objectivo de melhorar quer os modelos das Línguas, que o modelo de tradução.

10 http://www2.lael.pucsp.br/corpora/alinhador/
11 http://linguateca.di.uminho.pt/natools/
12 Statistical Machine Translation

Vejamos alguns exemplos:

1. Modelos de tradução baseados em palavras

Os sistemas actuais baseados na equivalência entre palavras são tipicamente extensões aos modelos propostos pela IBM (Brown, 1990), com o objectivo de melhorar o seu desempenho. Entre algumas das extensões propostas inclui-se a utilização de informação complementar para cada palavra, incluindo etiquetas morfo-sintácticas, de forma a limitar (ou permitir) a reordenação das palavras.

2. Modelos de tradução baseados em segmentos

Nestes modelos, cuja relevância e desempenho tem vindo a aumentar nos últimos anos, a tradução é efectuada em grupos de palavras ou segmentos (Marcu, 2002). O sistema "open-source" Moses (Koehn 2007a) é um exemplo de uma ferramenta de tradução automática baseada em segmentos. Note-se que os segmentos podem ter (ou não) uma relação directa com as estruturas sintácticas das frases. No caso afirmativo (Wang, 2007), a importância de existirem ferramentas computacionais com a capacidade de obter estruturas sintácticas é fundamental. Outra abordagem possível à segmentação é a utilização das entidades nomeadas como identificador de segmentos com uma forte coerência interna e que devem ser analisados de uma forma atómica (Koehn, 2003).

Koehn (2007b; 2010) propôs uma extensão ao modelo baseado em segmentos, de forma a integrar informação linguística (ou outra) a nível das palavras – etiquetas morfo-sintácticas, lema, informação morfológica – tendo obtido melhorias nos resultados finais.

3. Modelos de tradução baseados em gramáticas sintácticas

A utilização de gramáticas, que incorporam conhecimento da Língua através do recurso a regras sintácticas, tem sido efectuada por diversos investigadores, com resultados bastante positivos (Wu, 1998; Yamada, 2001; Melamed 2004). As gramáticas de dependências também foram objecto de análise e aplicação neste domínio (Melamed, 2003).

4. Modelos da Língua “destino”

Para resolver ambiguidades no processo de tradução é fundamental modelar a Língua “destino”. Este processo de modelação é tipicamente baseado na construção de modelos de “n-gramas” e no cálculo de probabilidades associadas à sequência de termos. No entanto, a incorporação de informação linguística mais profunda nestes modelos tem vindo a sofrer um incremento, através do recurso a modelos sintácticos (Wu, 1998; Marcu, 2006).

5. Processo de tradução (*decoding*)

O processo de tradução propriamente dito (*decoding*) é, em termos gerais, equivalente a um processo de análise de frases na Língua “origem” (*parsing*), sendo o resultado “lido” na árvore obtida. No entanto, e devido à existência de ambiguidades na Língua Natural e na sua representação, é possível que existam várias traduções possíveis para a mesma frase “origem”. Nesta situação, é efectuada uma ordenação das propostas de tradução obtidas, com base no modelo da Língua “destino”. Este modelo pode ser mais complexo e abrangente do que o que é utilizado no processo de tradução, incorporando modelos “n-gramas” e analisadores morfológicos e sintácticos. Uma abordagem alternativa é a utilização de Support-Vector Machines, tendo como *input* um conjunto de informação associada a cada frase (incluindo informação linguística), para classificar se as frases obtidas pertencem ou não à Língua “destino”.

6. Modelos semânticos

A utilização de informação semântica no processo de tradução tem vindo a ser objecto de alguma atenção nos últimos anos, sendo uma área bastante promissora. No entanto, e devido à dificuldade em se realizarem análises semânticas dos textos, as abordagens existentes ainda são bastante preliminares, recorrendo tipicamente a informação semântica lexical (Chan, 2007) e não a uma análise semântica profunda.

CONCLUSÕES E TRABALHO FUTURO

Neste artigo foi analisada a importância do uso de informação linguística em modelos de tradução automática baseados em aprendizagem.

Em concreto, apresentou-se a arquitectura geral dos tradutores automáticos estatítisticos e discutiram-se as abordagens existentes e a importância da informação linguística nessas abordagens. Identificaram-se algumas das ferramentas computacionais existentes para a Língua Portuguesa, com licenças de software-livre, que poderão permitir a construção de tradutores automáticos optimizados para a Língua Portuguesa.

Como trabalho futuro, é fundamental aplicar estas metodologias e ferramentas computacionais a corpora de textos em Língua Portuguesa, para os quais existam textos paralelos noutras Línguas, de forma a avaliar as abordagens propostas e permitir a criação de sistemas "abertos" de tradução automática de/para a Língua Portuguesa.

REFERÊNCIAS BIBLIOGRÁFICAS

Bick, Eckhard. 2000. *The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Famework*, Aarhus University Press.

Bos, Johan. 2008. Wide-Coverage Semantic Analysis with Boxer. In: *J. Bos, R. Delmonte (eds): Semantics in Text Processing. STEP 2008 Conference Proceedings. Research in Computational Semantics.* College Publications. p. .277-286.

Brown, P. F., Cocke, J., Pietra, S. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., Roossin, P. S. 1990. A statistical approach to machine translation. *Computational Linguistics.* 16, 2 (Jun), p. 79–85.

Cardoso, Nuno. 2008. REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto. *Encontro do Segundo HAREM, PROPOR.* Aveiro. Portugal.

Chan, Y. S., Ng, H. T., and Chiang, D. 2007. Word sense disambiguation improves statistical machine translation. In *Proc. of ACL*. 33–40.

Gonçalves, Teresa; Quaresma, Paulo. 2008. Text classification using tree kernels and linguistic information. In: M. Arif Wani, Xue wen Chen, David Casasent, Lukasz Kurgan, Tony Hu, Khalid Hafeez (Eds). *IMLA'08 – 7th International Conference on Machine Learning and Applications*. IEEE Computer Society. p. 763–768.

Gonçalves, Teresa; Quaresma, Paulo. 2009. Using graph-kernels to represent semantic information in text classification. In: Petra Perner (Ed). *Machine Learning and Data Mining in Pattern Recognition*, *Lecture Notes in Artificial Intelligence LNCS/LNAI*. Vol. 5632. Springer. p. 632–646.

Gonçalves, Teresa; Quaresma, Paulo. 2010. Using linguistic information and Machine Learning Techniques to Identify Entities from Juridical Documents, In: E. Francesconi; E. Montemagni; W. Peters; D. Tiscornia (Eds). *Semantic Processing of Legal Texts*, *Lecture Notes in Artificial Intelligence LNAI 6036*, Springer, p. 44-59.

Koehn, Philipp; Haddow, Barry; Williams, Philip; Hoang, Hieu. 2010. More Linguistic Annotation for Statistical Machine Translation. *Fifth Workshop on Statistical Machine Translation and Metrics MATR.*

Koehn, Philipp; Hoang, Hieu; Birch, Alexandra; Callison-Burch, Chris; Federico, Marcello; Bertoldi, Nicola; Cowan, Brooke; Shen, Wade; Moran, Christine ; Zens, Richard; Dyer, Chris; Bojar, Ondrej; Constantin, Alexandra; Herbst, Evan. Moses: Open Source Toolkit for Statistical Machine Translation. *ACL 2007*. Demo.

Koehn, Philipp; Knight, Kevin. 2003. Feature-Rich Statistical Translation of Noun Phrases. *ACL*.

Koehn, Philipp; Hoang, Hieu. 2007b. Factored Translation Models, *EMNLP.*

Leal, Ana Luísa Varani. 2009. AUTEMA-DIS - Arquitetura Computacional para Identificação da Temática Discursiva em Textos em Língua Portuguesa. Tese de Doutoramento. Universidade de Évora. Portugal. 217 p.

Lopez, Adam. 2008. Statistical Machine Translation. ACM Computing Surveys. Volume 40, nº 3.

Marcu, D.; Wong, W. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proc. of EMNLP*. 133–139.

Marcu, D., Wang, W., Echihabi, A., Knight, K. 2006. SPMT: Statistical machine translation with syntactified target language phrases. In *Proc. of EMNLP*. 44–52.

Melamed, I. D. 2003. Multitext grammars and synchronous parsers. In *Proc. of HLT-NAACL*. 79–86.

Melamed, I. D. 2004. Algorithms for syntax-aware statistical machine translation. In Proc. Of *TMI.*

Oliveira, Francisco; Wong, Fai; Hong, Iok-Sai; Dong, Ming-Chui. 2010. Parsing Extended Constraint Synchronous Grammar in Chinese-Portuguese Machine Translation. *The International Conference on Computational Processing of Portuguese, former Workshop on Computational Processing of the Portuguese Language (PROPOR).* Porto Alegre. Brasil.

Schimd, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *International Conference on New Methods in Language Processing*. Manchester. UK.

Simões, Alberto Manuel; Almeida, José João. 2001. jspell.pm — um módulo de análise morfológica para uso em processamento de linguagem natural. *In Actas da Associação Portuguesa de Linguística,* p. 485–495

Wang, C., Collins, M., and Koehn, P. 2007. Chinese syntactic reordering for statistical machine translation. In *Proc. of EMNLP-CoNLL*. 737–745.

Wu, D.; Wong, H. 1998. Machine translation with a stochastic grammatical channel. In *Proc. of ACL-COLING*. 1408–1415.

Yamada, K.; Knight, K. 2001. A syntax-based statistical translation model. In *Proc. Of ACL-EACL.*