POS tagging: clarificação histórico-terminológica

Diana Santos

Linguateca

www.linguateca.pt

A confusão do costume

- A maior parte das pessoas não tem rigor nos termos que usa nem clarifica suficientemente os conceitos
- *POS tagging* é um bom exemplo, e o resultado é que ninguém sabe muito bem o que é que se pretende designar

- Várias maneiras (não exclusivas) de definir um termo
  - Através do processo usado
  - Através do resultado obtido
- Algumas pessoas usam *POS tagging* para designar o resultado, outras o processo

O que é POS?

- POS: *part of speech* é a anglificação do termo latino *pars orationis*
- Em português tal é geralmente designado por classificação gramatical de uma palavra
- A classificação gramatical não é algo neutro e óbvio sobre o qual não pesam dúvidas. Muito pelo contrário,
  - Há diferenças importantes entre linguistas sobre esse tema
  - Há diferenças importantes entre as línguas
  - Há muitos casos que são vagos (podem ser igualmente classificados por mais de uma categoria)
- Um exemplo: nomes próprios são um subconjunto dos substantivos, ou uma classe à parte? (Diferença entre Portugal e Brasil)

O que é *tagging*?

- No processamento de linguagem natural / linguística computacional também não há um consenso sobre este tipo de terminologia
- Pode significar: associar uma marca ou etiqueta (*tag*)
- Costuma significar: associar uma marca a todas as unidades (*tokens*) presentes num texto, por oposição a *parsing* em que mais do que uma marca por palavra é esperada
- Pode significar:
  - uso de um sistema que assume que **uma** marca é suficiente
  - uso de um sistema que não usa outra informação do que as sequências de marcas possíveis
  - uso de um sistema fácil/rápido antes de uma análise mais complicada

O que é *POS tagging*?

- A atribuição de uma marca (*tag*) correspondendo à categoria gramatical de cada unidade num texto
- Que categorias gramaticais? O chamado *POS tagset*
- Como é que isso é feito, ou seja, usando que informação?
  - análise sintáctica completa?
  - análise sintáctica e semântica?
  - probabilidades de ocorrência das marcas em texto semelhante

Alguns exemplos completamente diferentes

- O PALAVRAS
- O Palavroso
- Os “multitaggers”

- Um radicalizador?

Mas vamos começar pela história do conceito...

Alguma história (1): os pioneiros russos

Nicolaeva (1958): análise sintáctica do russo para tradução automática

- *In order to obtain the desired dictionary information, it is necessary to reduce the word being analyzed to the form in which it may be found in the dictionary*
- *... The words are handled in accordance with their endings*
- *... Until the word is found*
- *... And find the following contextual and grammatical features of words with non-homonymic inflections*
- *The contextual features of words with homonymic inflections is ascertained on the basis of tactic and syntactic principles of context analysis*
- *Such an analysis requires larger semantically self-contained units: Phrase analysis*

Os pioneiros russos: Nicolaeva contd.

- Conforme os sinais de pontuação, definem sintagmas
- Usando toda a espécie de regras e com base na informação já sabida, tentam produzir mais informação morfológica (voz, caso, género)
- Depois vão para o que chamam sintaxe: *The sentence should constitute the unit of investigation, even though a study of word interconnections in a phrase is of great interest*
- *During the analysis of adjectives the presence of nouns syntagmatically related to the given ajectives is revealed. The adjective then receives the sign "agreeing attribute"... If not, instrumental case, ... If the number... part of compound predicate*
- *Nouns to be analysed are divided into 2 large groups: preceding prep.*

História: grammatical coding of English words

- **Klein & Simmons (1962):** *first component in a syntactic analysis program, which is part of a larger QA system*
- *An alternative to the use of a very large dictionary and can also be used as a context analyser which eliminates many ambiguities of word classes arising from the consideration of words in isolation*
- *The primary advantage of computation is that it avoids the labor of constructing a very large dictionary and permits a system to encode words it has never before encountered*
- *Most changes need be made only in dictionary and table entries (without modifying the running program)*
- *Tagging sentences in a computer-stored written English text with grammatical information permits a variety of data manipulations...*

Klein e Simmons (1962): CGC explicado

- Esclarecem que os nomes das classes foram escolhidos para facilidade de comunicação, mas que o número de classes depende da análise de cada um. Escolheram 30, mas o sistema genérico permite centenas
- *Because the analysis may be one of function, a form normally considered an adjective, for example, will occasionally be tagged NOUN because it functions as a noun* He chooses the **red**. He chooses the **beautiful**.
- Dicionários: de “function words”, de excepções para “content words”
- O resultado final pode ter mais do que uma análise!
- *Context Frame Test*: dada uma tríade em que a (série de) palavras do meio tem mais do que uma clasificação e as dos extremos são conhecidas, tentar determinar a classificação do meio

Exemplos de regras contextuais

- ARTICLE ADJ/VERB NOUN/ADJ VERB
- Indo à tabela, as possibilidades entre ARTICLE e VERB eram
  - ADJ NOUN
  - NOUN ADVERB
  - NOUN NOUN
- Para casos sem informação nenhuma, ficam NOUN/ADJ/VERB
- A tabela tem 500 entradas e no máximo sequências de 3 palavras ambíguas, em teoria poderia ter 2700 no meio (30x3x30) mas não são suficientemente frequentes (a partir de uma amostra de *Golden Book Encyclopedia text*)
- Casos em que não há solução são marcados com NONE, para inspecção subsequente do programa/tabela/dicionário/texto

Alguma história (3)

- Métodos estatísticos: Stolz et al. (1965)
- *Allocating each word of a language corpus to its respective grammatical form class or part of speech*
- *decisions ... based on conditional probabilities of various form classes in given syntactic environments*
- WISSYN produz apenas a classe mais provável, a partir de probabilidades estimadas a partir de um corpo de 30 mil palavras
- 18 classes, baseadas em Roberts, *Patterns of English*, 1956. Duas classes são de pontuação; *have_to*, *ought_to*, *should n't.* Apenas quatro classes são sujeitas à fase da probabilidade
- 60-70% das palavras são classificadas por simples consulta ao dicion., 10% pela fase adhoc (prep ou adv, auxiliares, gerúndios)

Ênfase no resultado: a anotação do *Brown corpus*

- Greene & Rubin (1971) aplicaram o algoritmo e a ideia de Klein & Simmons para criar o primeiro corpo anotado, mas usaram as classes do *Brown-Tougaloo English language project*
- Assumiram uma gramática do inglês que só atribui uma etiqueta certa
- Introduziram a noção de "internally ambiguous tag"
- Adicionaram "inflectional appendages" às etiquetas (plural, tempo, possessivo, contracções)
- Criaram um dicionário baseado em regras do estilo:
  - *a word ending in "-ing" could be tagged NN in addition to VBG if the construction "a ... was" was acceptable.* "a hearing was", "a skiing was"
  - *words of the form "un ... ed" where the base UN... is not a verb (e.g.* undecided*) are tagged JJ*

Como o Brown Corpus foi anotado

- Em duas fases: 1) todas as possibilidades; 2) depois, desambiguação, através de Context Frame Tests
- 900 frases do BC foram desambiguadas manualmente, e um programa criou todas as regras (CFT) necessárias para passar para o estado final
- Verificaram que sequências de duas ou três ambiguidades eram raras, passaram a fazer regras só com um elemento ambíguo, mas com dois elementos não ambíguos em qualquer dos lados
- Introduziram regras que podiam errar, desde que não errassem mais de 5% das vezes, mas cujo resultado estava marcado no fim
- As regras são aplicadas das mais específicas para as mais gerais
- E, no fim, edição manual: *manual post-editing phase*

Alguma história: aproximação de engenheiro

- Cherry (1978): *English might not lend itself to statistical methods... assign part of speech by rule*
- PARTS são 3 programas, o terceiro atribui os que os dois primeiros não conseguem, e corrige algumas das atribuições anteriores baseado no contexto
- Tem 20 classes internas, mas no fim só classifica em 10. As classes internas são de palavras ambíguas
- Conseguiu passar de 90% → 95%

Alguma história: o LOB corpus

- Marshall (1983) apresenta a anotação do LOB com base no Brown e com o objectivo de maior proporção de etiquetas correctas automaticamente
- Análise estatística das etiquetas no Brown Corpus produziu uma matriz de transições na qual o programa é baseado
- Algoritmo: calcular a sequência de etiquetas mais provável, entre duas etiquetas únicas
- Com base num sistema de limiares, se as probabilidades forem próximas a etiqueta tem de ser escolhida por seres humanos: o programa ordena pela probabilidade
- EVP ou expressões idiomáticas são identificados antes

Alguma história: A integração entre POS e análise sintáctica

- DeRose (1988) ou Church (1988) ou Garside et al. (1987) ou Hindle (1989) foram os primeiros que relataram a ligação entre um analisador sintáctico tradicional (baseado em regras e conhecimento linguístico) e um pré-processador para desbastar a floresta de derivações

- POS tagger como uma resposta de engenharia ao problema da explosão devida à ambiguidade
- Hindle (1989) fala de “lexical disambiguation”
- No seu analisador sintáctico determinístico, as regras aplicam-se a (46) categorias desambiguadas

Alguma história: primeira avaliação linguística?

- Macklovitch (1992): Primeira (?) análise linguística do POS tagging
- Fala de dependências globais (em vez de “long-distance”): se um dado verbo está no imperativo, no presente ou no conjutivo pode depender de toda a frase...
- E explica: *why bother*?
  - Pertinência da avaliação
  - Correcção ou pelo menos detecção automática de erros

O analisador de Brill (1992)

- "A simple Rule-Based PoS Tagger": *robust and rules automatically acquired*
- O que nós chamaríamos agora um **método híbrido**, baseado em AA
- A ideia: o próprio programa identifica e remedeia as suas fraquezas
- Primeiro, associa a etiqueta mais frequente às palavras que já existem em material de treino; depois usa as terminações das palavras OOD
- Depois compara-se o resultado da anotação com um corpo anotado e obtêm-se triplos de erros: <cat. antiga, cat. nova, número de vezes>
- Aplica-se um conjunto de 8 diferentes remendos, vendo qual o que dá a maior redução de erro global, e esse é aplicado à lista dos remendos
- 71 patches, 5% error in 5% of the Brown corpus

Alguma história, agora para o português

- `Encontra`, uma das ferramentas de corpos criadas no INESC baseadas no Palavroso (Medeiros, 1992), permitia procuras arbitrariamente complexas baseadas na morfologia (Palavroso = tal qual o sistema descrito em Cherry, mas desconhecido para nós)
- Medeiros et al. (1993) e o Palavroso: a primeira medição para o português da dificuldade da tarefa -- Resposta ou paralelo para o português: qual a ambiguidade existente e que precisa de ser resolvida? Quais os casos e sua frequência para o português
- Complementar: Bacelar do Nascimento el al. (1993) Ambiguidade morfológica no português fundamental

Na Linguateca

- A pouca pertinência do POS tagging for imediatamente reconhecida devido à existência do PALAVRAS (Bick, 2000): “tagging as parsing” e explicado
  - No Porto em 2000
  - Nos artigos Santos (1999) e Santos & Gasperin (2002)
- O que não impediu infelizmente que as pessoas continuassem a pedir financiamento ou a falar de *POS tagging* como uma tarefa relevante e interessante para o português, e internacionalmente
- Também não impediu a confusão que reina até hoje sobre corpos anotados (e revistos) e procura automática (através de gramáticas) em corpos
  - Ver artigo Santos & Ranchhod (1999)

Conclusão

Três tarefas diferentes:

- Análise morfológica: sem contexto, inerente à forma em si – o que não significa que não se possa ir longe!
- Análise morfossintáctica: devido ao contexto, o que podemos concluir
- Análise sintáctica: dependente da análise morfossintáctica, ou interdependente?
- Visão muito pessoal: nós precisamos de vários níveis de análise para "deitar fora o lixo", não pela análise em si:
  - fora/fora, como/como, cerca/cerca, uma/uma, era/era
  - fora: ser/ir, canto: da sereira ou da sala; baixo: voz ou tamanho; verde: cor ou maturidade

Conclusão (cont.)

- O mais importante é saber para que se quer a anotação morfossintáctica, e prosseguir em conformidade
- O mais importante é fazer avaliação linguística: que casos são importantes, que casos são irrelevantes?
- É sempre preciso estudo e trabalho da parte da pessoa que está a resolver o problema:
  - Ferramentas gerais precisam sempre de adaptação
  - Ferramentas específicas precisam sempre de melhoria e extensão

Conclusão final: porque estamos aqui?

- É importante ter uma tarefa concreta e bem definida para resolver
  - Gaizauskas (1998) *User-transparent* vs. ***User-visible***
- Áreas como *POS tagging*, análise morfológica, análise sintáctica, ou REM são apenas meios para outros fins (são transparentes para os U)
- Mas também não devemos idolatrar a tarefa: *That's Nice... What Can you Do With It?* (Belz, 2009)

- Um bom trabalho em PLN e LC ou EL contribui para a linguística, para a engenharia, para a informática, e para a realidade que nos cerca

Referências

- Maria Fernanda Bacelar do Nascimento, José Bettencourt Gonçalves, Lucília Chacoto, Paula Neto & Luísa Alice Santos Pereira. "Ambiguidade morfológica no Português Fundamental". In *Actas do 1º Encontro de Processamento da Língua Portuguesa (escrita e falada) (EPLP'93)* (Lisboa, 25-26 de Fevereiro de 1993), pp. 101-106.
- Belz, Anja. “That's nice ... what can you do with it?”, Last Words, *Computational Linguistics*, 35:1, pp. 111-118
- Bick, Eckhard. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press, 2000.

Referências (cont.)

- Brants, Thorsten. "Part of Speech Tagging". *Encyclopedia of Language and Linguistics*, 2nd Ed., ed. by Keith Brown. Oxford: Elsevier., 2006
- Brill, Eric. "A simple rule-based part of speech tagger", *Proceedings of the Third Conference on Applied Natural Language Processing* (Trento, Italy), 1992, pp.152-5.
- Cherry, Lorinda L. "PARTS - A System for Assigning Word Classes to English Text," Computer Science Technical Report #81, Bell Lab., Murray Hill, N.J., 1978.
- Church, Kenneth Ward. "A stochastic Parts Program and Noun Phrase Parser for Unrestricted Text", *Proceedings of the Second Conference on Applied Natural Language Processing*, (ACL), 1988, pp.136-43.

Referências (cont.)

- DeRose, Stephen J. "Grammatical category disambiguation by statistical optimization”. *Computational Linguistics* **14**, 1 (Jan. 1988), pp. 31-39.
- Garside, Roger, Geoffrey Leech & Geoffrey Sampson. *The Computational Analysis of English: A Corpus-Based Approach*, Longman, 1987.
- Gaizauskas, Robert. "Evaluation in language and speech technology". *Computer Speech and Language*, **12** (4) (1998), pp.249-62.

Referências (cont.)

- Greene, Barbara B. & Gerald M. Rubin. "Automated Grammatical Tagging of English". Providence, R.I.: Department of Linguistics, Brown University, 1971.
- Hindle, Donald. "Acquiring Disambiguation Rules from Text". ACL 1989, pp. 118-125
- Klein, Sheldon & Robert F. Simmons. "A computational approach to grammatical coding of English words". *Journal of the Association for Computing Machinery* **10**: 334-347.
- Macklovitch, Elliott. "Where the Tagger Falters", *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation* (Montréal, June 25-27, 1992), pp.113-26.

Referências (cont.)

- Marshall, I. "Choice of grammatical word-class without global syntactic analysis: Tagging words in the LOB Corpus". *Computers in the Humanities* **17** (1983), pp. 139-150.
- Medeiros, José Carlos. "Ferramentas de processamento de corpora usando o PALAVROSO", in Santos, Diana (ed), *Processamento de corpora no INESC*, Vol. 1, INESC Report RT-65/92, 1992.
- Medeiros, José Carlos, Rui Marques & Diana Santos: 1993, "Português Quantitativo", *Actas do 1.o Encontro de Processamento de Língua Portuguesa (Escrita e Falada) - EPLP'93* (Lisboa, 25-26 February 1993), pp. 33-8.

Referências (cont.)

- Nicolaeva, T. M. "Soviet Developments in Machine Translation: Russian Sentence Analysis", *Mechanical Translation* **5**, 2, November 1958, pp. 51-59.
- Santos, Diana. "Toward Language-specific Applications", *Machine Translation* **14** (2), June 1999, pp.83-112.
- Santos, Diana & Caroline Gasperin. "Evaluation of parsed corpora: experiments in user-transparent and user-visible evaluation", in Manuel González Rodríguez & Carmen Paz Suárez Araujo (eds.), *Proceedings of LREC 2002*, pp.597-604.

Referências (cont.)

- Santos, Diana & Elisabete Ranchhod. "Ambientes de processamento de corpora em português: Comparação entre dois sistemas", *Actas do IV Encontro sobre o Processamento Computacional da Língua Portuguesa (Escrita e Falada), PROPOR* (Évora, 20-21 de Setembro 1999), pp. 257-268.
- Stolz, Walter S., Percy H. Tannenbaum & Frederick V. Carstensen. "A stochastic approach to the grammatical coding of English". In *Communications of the ACM* **8** (6), June 1965, pp. 399-405.