X Congresso Brasileiro de Inteligência Computacional (CBIC'2011), 8 a 11 de Novembro de 2011, Fortaleza, Ceará

Geração de Sentenças em Português baseada na Teoria da Confabulação

Amaury Kruel Budri, Ricardo Suyama, Margarethe Born Steinberger-Elias
Universidade Federal do ABC - CECS/UFABC
{amaury.budri,ricardo.suyama, mborn}@ufabc.edu.br

Resumo – O presente artigo apresenta uma metodologia para geração automática de sentenças na língua portuguesa a partir de um fragmento de texto inicial. A abordagem é baseada na teoria da confabulação, que busca mimetizar a forma como o cérebro humano processa a informação. Os resultados obtidos com uma implementação simples e direta da teoria da confabulação, bem como os resultados obtidos com algumas modificações introduzidas nesse artigo, comprovam a viabilidade do método.

Palavras-chave – Predição de palavras, Teoria da Confabulação, Cogência, Processamento de Linguagem Natural.

Abstract – This paper presents a new methodology for sentence completion for the portuguese language. The proposed approach is based on the confabulation theory, which was developed to model how the human brain process information. The results obtained with the direct implementation of confabulation theory, and afterwards with a slightly modified scheme, attest the viability of the proposed method.

Keywords – Word prediction, Confabulation Theory, Cogency, Natural Language Processing.

1. Introdução

Um dos objetivos da área de pesquisa conhecida como *processamento de linguagens naturais* é estudar métodos computacionais para a geração automática e compreensão da linguagem humana [1], que distingue-se das demais formas de comunicação por codificar a informação sob a forma de cadeias de palavras. Por um lado, essas cadeias são agrupadas segundo restrições gramaticais e semânticas que permitem gerar padrões. Por outro lado, cada língua caracteriza-se por um grau de flexibilidade maior ou menor com respeito à ordem em que as palavras são encadeadas para compor sentenças e textos. Línguas como o Português e o Espanhol têm padrões de ordenação mais flexíveis do que o Inglês ou o Alemão.

Neste trabalho, consideramos as palavras como unidade elementar de informação lingüística e analisamos a possibilidade de predição de unidades seqüenciadas na composição de sentenças. Esse problema está diretamente relacionado a uma uma tarefa cognitiva relativamente simples para os seres humanos: a de compreender um determinado trecho de texto incompleto e propor palavras que completem a sentença, de maneira que a frase seja coerente.

Para realizar tal tarefa, contamos não somente com um vocabulário apropriado, construído ao longo dos anos, mas também nos baseamos em nossa habilidade de análise estrutural e semântica do texto apresentado. Diferentes abordagens já foram propostas para mimetizar tal habilidade humana, buscando soluções fundamentadas em métodos estatísticos. Um dos pioneiros nesse sentido foi Zipf [2], que analisou a distribuição de freqüências em que palavras são usadas para constituir sentenças, formalizando através da chamada Lei de Zipf o fato de que as freqüências lexicais de qualquer corpus são organizadas em ordem decrescente, caracterizando uma correspondência (do tipo *power law*) entre freqüência e ranqueamento de cada palavra de um corpus. Shannon [3] também trabalhou em temas correlatos, buscando analisar a predizibilidade de sequência de letras na língua inglesa, em uma tentativa de prever qual a seria a palavra subsequente a um dado fragmento de texto.

O problema tratado por Zipf e Shannon deu origem a uma linha de pesquisa hoje conhecida como predição de texto, e as técnicas desenvolvidas têm sido empregadas com sucesso, por exemplo, em sistemas de comunicação assitiva, permitindo que pessoas com severas deficiências motoras e orais possam se expressar mais rapidamente [4, 5]. Nesse tipo de sistema, o usuário tem acesso a uma lista de possíveis palavras a serem inseridas no texto, reduzindo significativamente o esforço de digitação dos usuários. Um exemplo desse tipo de aplicação foi apresentado em [6], trabalho no qual os autores propõem um teclado interativo, no qual a sequência de teclas previamente pressionadas era utilizada para prever quais seriam letras subsequentes mais prováveis. Nesse contexto, diversos outros trabalhos podem ser citados, como [7–10].

No presente trabalho, propomos uma metodologia para o problema geral de geração automática de sentenças, cujo objetivo é produzir sentenças completas, e não somente determinar a palavra subsequente, a partir de um fragmento de texto inicial. Para isso, adotamos um modelo de inferência baseado na *teoria da confabulação* [11], uma proposta que busca modelar a forma como os seres humanos processam a informação. O uso da teoria da confabulação no problema de previsão de sentenças foi inicialmente abordado em [11], mas até o momento apenas o caso específico da língua inglesa foi analisado [12].

A fim de expor nossa metodologia, o artigo foi estruturado da seguinte maneira. Inicialmente apresentaremos a abordagem estatística clássica para o problema. Na Seção III discutimos a teoria da confabulação, e como ela foi utilizada para compor o sistema de predição. Na Seção IV descrevemos o sistema implementado e apresentamos os resultados obtidos na seção subsequente. Finalmente, na Seção VI apresentamos as conclusões e discutimos algumas perspectivas para trabalhos futuros.

2. Abordagem Estatística

A tarefa de predição de palavras pode ser compreendida como uma instância do problema estimação, e pode ser resumido da seguinte forma: encontrar a sequência de palavras $\hat{W}$ mais provável, dado que o conjunto de palavras A foi observado. Em termos matemáticos, o problema consiste no desenvolvimento de um estimador de *máxima a posteriori*, i.e.,

$$\hat{W} = \arg\max_{W} P(W|A) \tag{1}$$

onde $P(W|A)$ representa a probabilidade condicional de W dado que A foi observado. Utilizando a regra de Bayes, podemos obter

$$\hat{W} = \arg\max_{W} \frac{P(A|W)P(W)}{P(A)} \tag{2}$$

onde $P(A|W)$ é a função de verossimilhança, $P(W)$ corresponde à probabilidade *a priori* do conjunto de palavras W e $P(A)$ a probabilidade dde A. Uma vez que $P(A)$ é independente da escolha de W, podemos simplificar (2) e obter

$$\hat{W} = \arg\max_{W} P(A|W)P(W) \tag{3}$$

Suponha, portanto, que o objetivo é determinar qual é a palavra w_N que completará a sentença formada por um conjunto de $N-1$ palavras w_1, w_2, ..., w_{N-1}. De acordo com (3), é necessário conhecer a probabilidade condicional

$$P(w_N|w_1, w_2, \ldots, w_{N-1}) \tag{4}$$

o que pode se tornar uma tarefa inviável caso o número de parâmetros for muito grande. Dessa forma, modelos simplificados foram propostos na literatura, dentre eles destaca-se o modelo de *n-gramas* [13], que assume que apenas um conjunto limitado de palavras afeta a probabilidade da palavra subsequente. Em outras palavras,

$$P(w_i|w_1, w_2, \ldots, w_{i-1}) \approx P(w_i|w_{i-n+1}, w_{i-n+2}, \ldots, w_{i-1}) \tag{5}$$

O modelo de n-gramas é amplamente difundido em processamento de linguagem natural, e uma simplificação adicional é obtida ao considerar que as palavras de uma sentença dependem apenas da palavra imediatamente anterior a ela, i.e., considerar apenas *bigramas* ($n = 2$). Nesse caso, o trabalho de levantamento das probabilidades fica drasticamente reduzido, resumindo-se à estimação de $P(w_i|w_{i-1})$ para cada uma das palavras encontradas no corpus de treinamento.

A equação (3) deixa claro que o estimador leva em consideração a probabilidade *a priori* de uma determinada palavra (ou conjunto de palavras) aparecer no texto. Esse aspecto, embora seja bastante relevante em outros contextos, como no caso de classificação de padrões [14], pode levar a sugestões enviesadas no caso de previsão de palavras em uma sentença [11]. Tome, por exemplo, o caso da lingua inglesa, na qual a palavra `the` é uma das mais frequentes. Nesse caso, a probabilidade a priori da palavra pode fazer com que a decisão final recaia sobre a palavra `the`, muito embora outras possibilidades, que resultariam em um texto semânticamente mais rico, pudessem ser inseridas.

3. Teoria da Confabulação

A teoria da confabulação, proposta por Hetch-Nielsen, se baseia na premissa de que todos os aspectos da cognição - visão, audição, pensamentos, planejamento, linguagem, pensamento abstrato etc. - podem ser explicados por meio da interação entre diferentes áreas do cérebro, cada uma delas responsável pela representação mental de um objeto.

Segundo a teoria proposta por Hetch-Nielsen, cada hemisfério do córtex cerebral humano pode ser dividido em, aproximadamente, 2000 *módulos funcionais*, que são controlados de maneira independentes. Cada um desses módulos seria responsável por representar um atributo de um objeto - seja ele visual, auditivo, abstrato etc. Essa representação se dá através de um subconjunto de neurônios, cuja ativação está associada à seleção de um único símbolo dentre um conjunto de milhares de símbolos implementados pelo módulo.

Nesse contexto, o conhecimento adquirido é representado por meio de conexões entre os símbolos armazenados, e o processo de apredizagem segue o princípio Hebiano: os elos entre dois símbolos são fortalecidos na medida em que os dois conjuntos de neurônios, cada um representando um símbolo, são ativos [11, 15].

A hipótese básica da teoria proposta por Hetch-Nielsen é que o processo cognitivo envolve apenas um tipo de processamento da informação: a *confabulação*, um tipo especializado de competição *winners-take-all* que ocorre entre os símbolos de um módulo, após terem recebido estímulos externos provenientes dos elos formados ao longo do processo de aprendizagem.

Um exemplo de como ocorre o processo de confabulação é ilustrado na Figura 1, onde 5 módulos estão envolvidos. Cada um dos módulos à esquerda descreve atributos de um ou mais objetos mentais, sendo que cada módulo expressa um único símbolo (no caso, α, β e γ). Cada um dos símbolos expressos possui um certo número de ligações com símbolos pertencentes ao quarto módulo. No caso ilustrado destacam-se apenas algumas das ligações entre os os símbolos α, β e γ, e os símbolos do quarto módulo.

Cada uma das ligações faz com que um determinada quantidade de estímulo externo seja aplicada a cada um dos símbolos do quinto módulo, e o processo de confabulação determina que o símbolo "vencedor" seja aquele com o maior nível de ativação dentro do módulo.

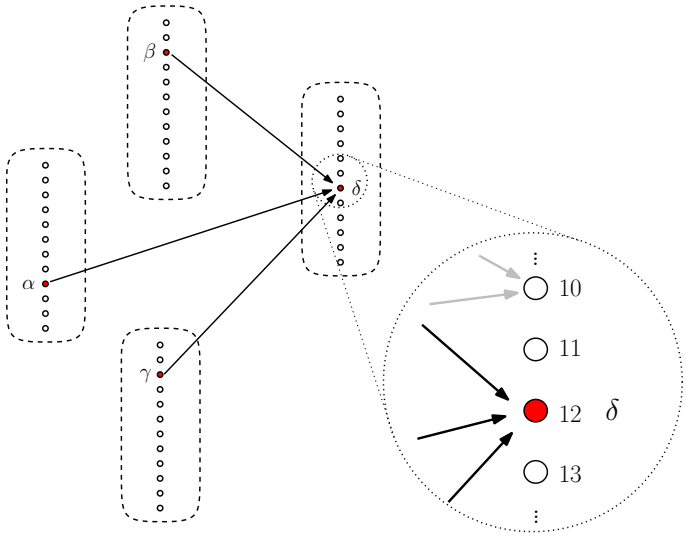

Figura 1: Exemplo de confabulação, considerando ligações de conhecimento entre 5 símbolos.

Em termos matemáticos, a resposta obtida pelo processo de confabulação corresponde a escolher o símbolo δ tal que a função

$$P(\alpha|\delta) \cdot P(\beta|\delta) \cdot P(\gamma|\delta) \tag{6}$$

seja maximizada. Nesse caso, cada uma das probabilidades condicionais presentes na expressão estão relacionadas à ligação existente entre os símbolos dos diferentes módulos e o quinto módulo. É interessante notar que, sob a hipótese de que os símbolos α, β e γ são independentes dado o símbolo δ, o produto em (6) representa a *função de verossimilhança* $P(\alpha, \beta, \gamma|\delta)$, que na teoria da confabulação recebe o nome de *cogência* [11].

4 Confabulação aplicada à geração automática de sentenças

Uma vez que os elementos básicos da teoria de confabulação tenham sido apresentados, podemos aplicá-los ao nosso problema específico. Considere, para isso, que cada um dos módulos existentes represente um certo número de símbolos, cada um representando uma palavra provenientes de um conjunto de textos em um idioma específico.

Nesse contexto, suponha que os símbolos α, β e γ tenham sido observados, i.e., sejam palavras de uma sentença que começou a ser escrita pelo usuário. O objetivo de nosso sistema deve ser, portanto, determinar qual o símbolo δ que seria o mais adequado para completar a sentença iniciada. De acordo com a teoria da confabulação, o resultado esperado seria aquele tal que a cogência $P(\alpha, \beta, \gamma|\delta)$ é maximizada. É importante ressaltar, no entanto, que, conforme apresentado em (6), a cogência pode ser aproximada por meio das probabilidades condicionais.

4.1 Confabulação Simples

Baseado nessa idéia, a implementação do processo de completamento de sentenças utilizando o mecanismo da confabulação pode ser feito a partir do seguinte procedimento:

1. É feita a leitura dos textos de um corpus e computada a frequência de ocorrência simultânea de duas palavras, i.e., são estimadas as probabilidades condicionais.

2. A partir desse levantamento inicial, é computada a frequência de palavras que aparecem juntas na frase, separadas por uma palavra, por duas palavras, e assim por diante, até separadas por um dado número máximo $n - 1$ de palavras.

3. Com base nas frequências computadas nos Passos 1 e 2, calculam-se as cogências correspondentes. Em outras palavras, dado um fragmento de texto observado, calcula-se qual seria o valor correspondente de (6) para cada uma das possíveis palavras a serem inseridas no texto.

4. Finalmente, levando-se em conta os valores da cogência, o completamento da frase é feito concatenando-se, recursivamente, palavras à frase. A palavra escolhida, é aquela que maximiza a cogência, levando em conta as n palavras anteriores. O completamento é terminado quando é concatenada uma palavra de pontuação (ponto final, interrogação, exclamação)[1], ou um número máximo de palavras tenha sido atingido.

[1]Caracteres de pontuação são tratados como palavras.

Tabela 1: Fragmentos de textos iniciais, utilizados nos testes realizados com o sistema de predição.

- O presidente Luiz Inácio
- Ele foi a Brasília
- Ele esteve em Brasília
- Este programa não está
- Falei que levaria um
- O suspeito foi levado
- A ministra Dilma disse
- Ele está em Brasília
- Os preços dos produtos
- Será assim até entrar
- O time não conseguiu
- Parlamentares em Brasília negociavam

Tabela 2: Sentenças obtidas a partir de um processo de confabulação simples palavra a palavra.

- O presidente Luiz Inácio Lula da Silva anunciou ontem , em São Paulo , que , a partir de junho , que das não permanentes , que são os de regularidade e segurança ; de : Da Guia) ; Douglas (Wellington Saci , Diego e Marcelo Oliveira Santos foram à área e trabalham para controlar o incêndio " , afirmou o promotor.
- A ministra Dilma disse " que não conhecia muito bem e não dá boas ferramentas para abrir um na 16 Volta às aulas :: 35 professores ::
- Ele foi a Brasília e já percebi um quarto da minha sala e não estava demais comigo , mas não tenho muita certeza é que gera uma pressão social elevada carga tributária que já assusta mais investimento que fizemos para ter infra - estrutura começará a avaliar o elenco durante a Copinha e o Paulista da Série A
- Ele está em Brasília e não a cair na água fica mais fácil se livrar do de Luis , suspenso pelo terceiro cartão amarelo , além dos volantes Márcio , ex -
- Os preços dos produtos agrícolas subiram 3 , 69 % em julho .
- Este programa não está sendo seguido pela polícia , ele , que deu um pênalti em um zagueiro Renan , do direito Alex Silva , que estava no Coritiba , e o Guaratinguetá , do ex - treinador do Rio Preto foi realizar , a partir de maio , cirurgias de médio porte , tanto no mar - de -
- Será assim até entrar , diz , não " nada satisfeito gera a , resulta em 30 % de economia .
- Falei que levaria um século para ser concedido às mulheres que fizeram uso do
- O time não conseguiu levar seu jogo de futebol , sobretudo quando o seu time , o Santos , está em campo com o mais restrito da polícia que , apenas ganhar dinheiro , poder tudo fazer qualquer coisa que não quero na vida nada na vida além de ter contato com toda a " que está sofrendo pressão na sua
- O suspeito foi levado por policiais e acusado de tentar matar P .
- Parlamentares em Brasília negociavam.

4.2 Confabulação hierárquica

Utilizando a metodologia descrita, e considerando os fragmentos de frases iniciais listados na Tabela 1, geramos automaticamente as frases apresentadas na Tabela 2. Conforme pode-se observar, a aplicação simples e imediata do processo de confabulação leva a resultados insatisfatórios. A tentativa de obter, uma a uma, a próxima palavra a ser incorporada na frase, a partir das palavras imediatamente anteriores produz sentenças mal formadas, com diversos erros gramaticais. A causa evidente é que a linguagem não emprega um encadeamento sequencial e linear de palavras, sendo, portanto, um processo cognitivo mais complexo. Para obter resultados melhores e mais realistas devemos incorporar um nível adicional de confabulação, realizando uma confabulação hierárquica.

A confabulação hierárquica não concatena palavras isoladas à frase. Ao invés de trabalhar apenas com as palavras, em um primeiro momento buscamos concatenar palavras isoladas a fim de formar frases curtas. Após essa etapa inicial, ocorre um segundo processo de confabulação que visa aglutinar frases para formar a sentença. Uma maneira simples de implementar a confabulação hierárquica se baseia na criação de símbolos que correspondam a sequências de palavras muito frequentes no texto. Por exemplo, na língua portuguesa, é muito frequente a ocorrência, por exemplo, das palavras `de que`. Dessa forma, cria-se um símbolo que represente duas palavras, e durante todo processamento descrito na Seção 4.1, o termo "de que" é processado como se fosse uma única palavra.

4.3 Formação de consenso

A confabulação hierárquica não é o único mecanismo através do qual podemos melhorar os resultados da confabulação simples. De fato, suponha que para uma dada frase inicial exista um conjunto de palavras candidatas a completar a frase, e o valor da cogência para cada uma das possibilidades é muito semelhante. Nesse caso, talvez seja mais interessante realizar um procedimento adicional, que denominaremos aqui de *formação de consenso*. Neste procedimento, em vez de se considerar uma

única palavra para maximização da cogência, são consideradas sequências de C palavras. Em outras palavras, seleciona-se a sequência de C palavras (ou menores, se terminadas por pontuação) que maximiza a cogência, e a primeira palavra da sequência é concatenada à frase inicial. O processo então se repete até ser concatenada uma palavra de pontuação.

5 Resultados

Com a finalidade verificar a metodologia apresentada anteriormente, foram realizados diferentes experimentos a partir de um corpus composto por textos jornalísticos escritos em língua portuguesa, contendo cerca de cinco milhões de palavras. Os experimentos, a exemplo dos resultados obtidos para a confabulação simples, consistiram na geração automática de sentenças a partir dos fragmentos de texto apresentados na Tabela 1.

Os resultados obtidos com a proposta de confabulação hierárquica são apresentados na Tabela 3. Neste caso, observa-se que muitas frases propostas são coerentes, e há uma ocorrência muito menor de violações gramaticais. Os mecanismos avaliados, no entanto, ainda estão distantes de um desempenho humano, mas é possível identificar e propor soluções para alguns dos problemas.

Um dos aspectos observados se refere à tendência do mecanismo de confabulação em ignorar palavras que são muito frequentes [16], uma vez que sua cogência é, em geral, menor do que a de palavras pouco frequentes. Podemos observar isso em frases como "A ministra Dilma disse presidente que não vai ganhar...". O mecanismo poderia ter proposto "disse ao presidente", no entanto o método de confabulação hierárquica falhou em agrupar as palavras "disse" e "ao" num único símbolo. Isto aponta para a necessidade o emprego de mecanismos hierárquicos mais sofisticados. Por outro lado, conforme comentado anteriormente, incluir a informação *a priori* a fim de obter uma estimativa MAP para a palavra subsequente poderia polarizar os resultados, também gerando sentenças sem sentido. Numa tentativa de se reduzir esse problema, modificamos o mecanismo de maximização da cogência, incluindo classes de palavras frequentes e pouco frequentes. Com isso, caso a palavra analisada pertença à classe de palavras frequentes, um fator multiplicado fixo foi incluído no cálculo da cogência. Desta forma, o mecanismo de maximização da cogência incluirá, eventualmente, palavras consideradas frequentes. Os resultados obtidos após a modificação são mostrados na Tabela 4, e é possível notar algumas diferenças nas construções das frases, principalmente no comprimento das sentenças, uma vez que os caracteres de pontuação que encerram as frases são tratados como palavras frequentes.

Tabela 3: Sentenças obtidas a partir de um processo de confabulação hierárquica e com formação de consenso.

- O presidente Luiz Inácio Lula da Silva sancionou ontem a lei que cria motovias nas principais avenidas da cidade .
- A ministra Dilma disse presidente que não vai ganhar com isso e fazer da vida .
- Ele foi a Brasília e já estar logo ao me enfrentar isso novamente algumas atitudes melhores valores são diferentes em cada .
- Ele está em Brasília , com é o caso de Vaz de Lima .
- Os preços dos produtos agrícolas subiram 3 , 69 % em julho , em comparação com a alta de 3 , 37 % e , no ano , de 53 , 67 % .
- Este programa não está sendo seguido pelo bancos e , não sendo uma " parada técnica muito antiga " , afirma .
- Será assim até entrar , diz .
- Falei que levaria um século para ser esse poder " .
- O time não conseguiu levar seu jogo de futebol , mesmo que seja muito melhores !
- O suspeito foi levado à delegacia da cidade , onde acabou autuado em flagrante pelo crime sexual .
- Parlamentares em Brasília negociavam Olegário , e Neuza Aparecida da Silva , outra da BM&F , e outra abstrata .

6 Conclusões

Nesse trabalho, testamos uma metodologia para geração de sentenças a partir de fragmentos iniciais de texto que se baseia na teoria da confabulação. Os resultados mostram que o mecanismo de confabulação é uma promissora técnica para processamento de linguagem natural. No entanto, ainda há margem para melhoramentos, incluindo mecanismos de confabulação hierárquicos mais complexos, além da incorporação de informação linguística como os graus de afinidade semântica entre termos e seu potencial de substituibilidade, para que este processamento obtenha melhores resultados.

Referências

[1] Z. S. Harris. *Mathematical Structures of Language*. John Wiley, New York, 1968.

[2] G. K. Zipf. *The psycho-biology of language*. Houghton Mifflin, Boston, 1935.

Tabela 4: Sentenças obtidas a partir de um processo de confabulação hierárquica, com formação de consenso e correção do fator de cogência para palavras frequentes.

- O presidente Luiz Inácio Lula da Silva afirmou ontem , em Madri .
- A ministra Dilma disse presidente que não vai ganhar com isso não .
- Ele foi a Brasília e já estar logo ao me ajuda muito disso porque vou dá - lo .
- Ele está em Brasília , com é o caso de Vaz de Lima .
- Os preços dos produtos agrícolas subiram 3 , 69 % em julho , em comparação com a alta de 3 , 37 % e , no ano , de 53 .
- Este programa não está sendo seguido pelo acusado .
- Será assim até entrar , diz .
- Falei que levaria um século .
- O time não conseguiu levar seu jogo de futebol , mesmo que seja muito melhores !
- O suspeito foi levado à delegacia da cidade , onde acabou autuado em flagrante pelo crime sexual .
- Parlamentares em Brasília negociavam o diretor - executivo da PF .

[3] C. E. Shannon. "Prediction and entropy of printed English". *Bell Systems Technical Journal*, vol. 30, pp. 50–64, 1951.

[4] N. Garay-Vitoria and J. Abascal. "A Comparison of Prediction Techniques to Enhance the Communication Rate". In *User-Centered Interaction Paradigms for Universal Access in the Information Society*, edited by C. Stary and C. Stephanidis, volume 3196 of *Lecture Notes in Computer Science*, pp. 400–417. Springer Berlin / Heidelberg, 2004.

[5] W. Zagler and C. Beck. "FASTY - faster typing for disabled persons". In *Proc. European Conference on Medical and Biological Engineering*, 2002.

[6] J. J. Darragh, I. H. Witten and M. L. James. "The Reactive Keyboard: A Predictive Typing Aid". *Computer*, vol. 23, pp. 41–49, 1990.

[7] H. Motoda and K. Yoshida. "Machine learning techniques to make computers easier to use". In *Proc. 15th International Joint Conference on Artificial Inteligence*, Nagoya, Japão, 1997.

[8] B. D. Davison and H. Hirsh. "Predicting Sequences of User Actions". In *Proc. AAAI/ICML Workshop on Predicting the Future: AI Approaches to Time Series Analysis*, pp. 5–12. AAAI Press, 1998.

[9] N. Jacobs and H. Blockeel. "User modeling with sequential data". In *Proc. 10th International Conference on HCI*, pp. 557–561, 2003.

[10] B. Korvemaker and R. Greiner. "Predicting UNIX Command Lines: Adjusting to User Patterns". In *Proc. 17th National Conference on Artificial Intelligence*, pp. 230–235. AAAI Press, 2000.

[11] R. Hecht-Nielsen. "Cogent confabulation". *Neural Networks*, vol. 18, pp. 111–115, March 2005.

[12] Q. Qiu, Q. Wu, D. Burns, M. J. Moore, R. E. Pino, M. Bishop and R. Linderman. "Confabulation Based Sentence Completion for Machine Reading". In *Proc.IEEE Symposium Series in Computational Intelligence 2011*, Paris, França,, 2011.

[13] C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, first edition, June 1999.

[14] R. O. Duda, P. E. Hart and D. G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, second edition, November 2001.

[15] R. Hecht-Nielsen. *Confabulation Theory: The Mechanism of Thought*. Springer, first edition, 2007.

[16] H. Sekiya, T. Kondo, M. Hashimoto and T. Takagi. "Context representation using word sequences extracted from a news corpus". *International Journal of ApproximateReasoning*, , no. 45, pp. 424–438, 2007.