Minimização do Impacto do Problema de Desvio de Conceito por Meio de Acoplamento em Ambiente de Aprendizado Sem Fim

Maisa Cristina Duarte, Estevam R. Hruschka Jr., Maria do Carmo Nicoletti

Departamento de Computação – Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676, 13565-905 – São Carlos – SP – Brazil

{maisa_duarte,estevam,carmo}@dc.ufscar.br

Abstract. *Machine Learning (ML) is a research subarea of Artificial Intelligence that aims to develop computer programs that can evolve with new experiences. Among the many ML goals, the endless learning, i.e., methods that would enable computer systems to autonomously improve their own performance, based on previously learnt information, is of particular interest in the research described in this paper. In the framework of endless learning this paper investigates semi-supervised ML techniques which would minimize the impact of a problem known as concept drift.*

Resumo. *Aprendizado de Máquina (AM) é uma subárea de pesquisa de Inteligência Artificial que objetiva o desenvolvimento de programas computacionais que podem evoluir à medida que vão sendo expostos a novas experiências. Entre os muitos objetivos de AM, o aprendizado sem fim, i.e., métodos que capacitam sistemas computacionais a melhorar seus próprios desempenhos, de maneira autônoma, usando informação previamente aprendida, é de interesse na pesquisa descrita neste artigo. Em um ambiente de aprendizado sem fim, este artigo investiga técnicas de AM semissupervisionadas que podem minimizar o impacto de um problema conhecido como desvio de conceito.*

1. Introdução

Desde o estabelecimento de Aprendizado de Máquina (AM) como uma área de pesquisa *per se*, o contínuo aumento no volume de publicações disponíveis na literatura, relacionadas às diferentes técnicas, formalismos e abordagens envolvendo aprendizado automático, é evidência que AM continua bastante ativa e produtiva. Apesar de todo o investimento em pesquisa ao longo dos vários anos, entretanto, o projeto e implementação de um sistema computacional totalmente automatizado, que aprenda de maneira incremental e que, também, seja capaz de usar o conhecimento previamente aprendido para, continuamente, ir refinando sua capacidade de aprendizado ao longo do tempo, ainda parece um objetivo razoavelmente distante de ser atingido. Um sistema que assim funcionasse poderia ser caracterizado como um "Sistema de Aprendizado Sem Fim" (SASF) i.e., um sistema que, de maneira autônoma, viabiliza sua constante e contínua evolução ao longo do tempo. Teoricamente em um SASF o aprendizado deve acontecer continuamente; o conhecimento adquirido servirá, também, para

Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pages 134–143,
Cuiabá, MT, Brazil, October 24–26, 2011.

dinamicamente realimentar o sistema fazendo com que o seu escopo se expanda e o seu desempenho melhore.

Por meio de uma análise cuidadosa das características de um SASF é possível notar que elas são similares às de sistemas computacionais que implementam o chamado Aprendizado Semissupervisionado (AS) [Chapelle et al. 2006]. Ambas as abordagens de aprendizado buscam aprender mais, a partir de um pequeno volume de conhecimento inicial, utilizando informações não categorizadas. Esta similaridade motivou parte da comunidade de AM à investigação e ao desenvolvimento de métodos de AS que incorporassem características específicas de aprendizado sem fim. O uso de técnicas de AS na geração de um SASF, entretanto, provoca um problema conhecido como "desvio do conceito" (*concept drift*). O desvio de conceito não significa necessariamente um erro. É fato, entretanto, que determinados conceitos podem mudar com passar do tempo e a identificação das mudanças é crucial para manter a corretude do conhecimento aprendido. O desvio de conceito ocorre em AS, por exemplo, quando o sistema categoriza, de maneira incorreta, um novo exemplo e, a partir daí, passa a utilizá-lo para categorizar novos exemplos (provavelmente provocando incorreções). Essa característica faz com que, com o passar do tempo, o sistema aprenda mais e mais conceitos incorretos, inviabilizando o seu uso [Curran et al. 2007].

A hipótese de pesquisa que norteou o trabalho descrito neste artigo é a de que o problema de desvio de conceito associado aos métodos de AS pode ser minimizado. Para tanto, deve-se acoplar tarefas de aprendizado de maneira a permitir que resultados de algoritmos de aprendizado semissupervsionado possam ser (automaticamente) utilizados pelo SASF para minimizar o desvio de conceito. A idéia foco foi, pois, o uso de acoplamento de tarefas de AS. Buscando evidências empíricas para dar suporte à hipótese de pesquisa, após o estudo e investigação de diferentes métodos e técnicas relacionados à construção de SASFs, foi desenvolvido o SASF chamado RTWP (*Read The Web in Portuguese*), para a extração de conhecimento de textos de páginas da Web em português. Os resultados obtidos e descritos neste artigo são evidência empírica que há uma tendência na melhoria da acurácia de aprendizado que faz uso de acoplamentos. O restante deste artigo está organizado como segue. A Seção 2 apresenta uma descrição mais detalhada do principal objetivo do trabalho de pesquisa realizado. A Seção 3 focaliza os principais passos da metodologia de trabalho adotada. A Seção 4 sumariza algumas propostas e resultados recentes e relevantes, relacionados ao trabalho desenvolvido. A Seção 5 descreve o sistema RTWP (*Read The Web in Portuguese*) e os acoplamentos considerados. Os experimentos e discussão dos resultados obtidos com o uso dos acoplamentos propostos são descritos na Seção 6 e, finalmente, na Seção 7, são apresentadas conclusões e perspectivas para a continuidade do trabalho desenvolvido.

2. Caracterização Detalhada do Objetivo Principal do Trabalho

O principal objetivo de pesquisa foi investigar, propor e implementar métodos e algoritmos de AS que permitam a minimização do problema de desvio de conceito. Para tanto, foi construído um sistema computacional capaz de realizar a extração de conhecimento a partir da Web em português, por meio da criação de uma base de conhecimento consistente, constantemente atualizada à medida que novos conhecimentos vão sendo extraídos por meio do AS. O sistema computacional proposto e implementado (RTWP) é capaz de, a partir de uma base de conhecimento inicial (representada em uma estrutura de ontologia), ser executado continuamente por um

determinado período de tempo, com dois objetivos específicos, a saber: *(1) Extração:* extrair mais "conhecimento" a partir da Web em português visando à expansão da base de conhecimento inicial; *(2) Aprendizado:* aprender a extrair melhor e com mais precisão que "antes".

A ontologia adotada, brevemente descrita na Seção 6, foi inspirada na utilizada em [Betteridge et al. 2009b]. Um exemplo de ontologia inicial pode ser visto em http://rtw.ml.cmu.edu/readtheWeb.html, que é a própria ontologia do sistema NELL, abordado na Seção 4 deste trabalho e do qual o RTWP é parte. Os algoritmos desenvolvidos e implementados foram avaliados com relação ao desempenho bem como à capacidade de continuar aprendendo com o passar do tempo. Assim, a avaliação foi realizada com foco nos objetivos (1) e (2) descritos anteriormente. Com base em avaliações realizadas por meio de inspeção humana, foi possível identificar quando o sistema conseguiu melhorar seu desempenho com o uso dos métodos de acoplamento propostos e implementados.

3. Metodologia de Trabalho

A ideia que fundamenta a metodologia de trabalho é que a integração de vários processos de AS [Zhu et al. 2009] promove a minimização da divergência do aprendizado refletida no problema do desvio de conceito. Apesar de alguns pontos em comum, a metodologia utilizada neste trabalho não está vinculada às ideias de extração de informação aberta, como tratadas em [Banko et al. 2007] [Etzioni et al. 2011] e tampouco ao uso de aprendizado incremental. Está solidamente ligada aos princípios de aprendizado sem fim (*never-ending learning*) [Betteridge et al. 2009a]. Resultados apresentados em [Carlson et al. 2010] sugerem que o aprendizado integrado de tarefas de AS pode viabilizar o desenvolvimento de SASFs e, empiricamente, sustentam a viabilidade do uso de tarefas/métodos integrados para o desenvolvimento de um sistema computacional que incorpora características do ASF para a Web em português. Presentemente as tarefas integradas no RTWP são:

- Identificação e extração a partir de páginas Web de "Entidades Nomeadas" (ENs) a partir de Padrões Textuais (PTs);
- Identificação e extração a partir de páginas Web de "Padrões Textuais" (PTs) a partir de ENs;
- Identificação e extração a partir de páginas Web de Pares de ENs a partir de PTs de Relações Semânticas;
- Identificação e extração a partir de páginas Web de PTs de Relações Semânticas a partir de Pares de ENs.

Por limitação de espaço e devido ao foco do trabalho, não é objetivo introduzir uma notação padrão e apresentar, de maneira formal, os conceitos usados. Por essa razão, informalmente, por ENs entende-se os substantivos encontrados (ver [Whitelaw et al. 2008]). PTs podem ser informalmente definidos como as palavras que estão antes ou após uma EN ou, ainda, que estão entre pares de ENs e, finalmente, por Relações Semânticas podem ser entendidas as relações entre ENs.

A integração das quatro tarefas brevemente descritas anteriormente seguiu a metodologia definida em [Betteridge et al. 2009b]. Assim, a identificação de ENs com base em PTs foi integrada (por meio de um mecanismo de *Bootstrapping* [Yarowsky

1995]) caracterizando um método acoplado de identificação de ENs e PTs. Da mesma forma, a identificação e extração de pares de ENs por meio de PTs de relações semânticas, foi acoplada à identificação e extração de PTs de relações semânticas a partir de pares de ENs. O processo de acoplamento é fundamentado na idéia que, quanto maior o seu número (desde que sejam acoplamentos adequados), maior será o ganho no desempenho do sistema que os utiliza. A Seção 5 aborda acoplamentos com mais detalhes.

4. Trabalhos Relevantes Relacionados

O RTWP foi inspirado no sistema NELL (*Never-Ending Language Learning*) [Betteridge et al. 2009a, 2009b], [Blum and Mitchell 1998] e [Carlson et al. 2010], presentemente em desenvolvimento pela Carnegie Mellon University. Toda a concepção do RTWP foi investigada e proposta considerando as características específicas da língua portuguesa. Da mesma forma, a implementação foi desenvolvida do início (não foram traduzidos e tampouco utilizados métodos já definidos e implementados para o NELL). Outro ponto que diferencia o RTWP do componente de extração de ENs a partir de texto implementado no NELL, é o fato do RTWP extrair conhecimento diretamente da Web, enquanto que no NELL a extração é feita a partir de um *corpus* de páginas Web pré-processado e armazenado em disco.

O NELL está sendo desenvolvido com a intenção de definir formalmente e comprovar que a técnica de "aprendizado sem fim" é eficiente e viável tanto na teoria quanto em aplicações reais. Pretende-se mostrar que um computador pode continuamente adquirir conhecimento e ter autonomia suficiente para revisar e ampliar seu conhecimento a partir de novas descobertas apenas com o uso de uma eventual supervisão pontual (quando necessário). Existem técnicas que fazem isso, porém a confiabilidade tende a cair com o passar das iterações. A referência [Betteridge et al. 2009a] apresenta uma definição formal do sistema NELL que pode ser abordada por meio dos seus seguintes objetivos principais: (1) Ser autônomo; (2) aprender 24h por dia para sempre; (3) possuir uma ontologia dinâmica, sempre atualizada e (4) ser auto-supervisionado. Em [Betteridge et al. 2009b] são mostrados os primeiros resultados obtidos com o NELL, evidenciando a viabilidade do projeto. A referência [Blum and Mitchell 1998] apresenta e discute o algoritmo *Co-training* que subsidia a técnica usada para os acoplamentos propostos. Em [Carlson 2010] são apresentados, de forma mais detalhada, os resultados do NELL bem como um histórico de seu desenvolvimento e em [Mitchell et al. 2009] a macroleitura, que se refere à forma de aprendizado usada tanto no NELL quanto no RTWP, é abordada em detalhes. Nesse sentido, a abordagem que subsidia o RTWP é mais voltada às concepções baseadas em *machine reading*, como proposta em [Etzioni et al. 2006], [Banko et al. 2007], [Banko and Etzioni 2008] e [Suchanek et al. 2007] do que àquelas com base em Processamento de Lingua Natural (PLN).

5. RTWP – Read The Web in Portuguese: Características e Funcionamento

Neste trabalho entende-se por extração de conhecimento a partir da Web a capacidade automática e autônoma de um sistema computacional de realizar a leitura de páginas da Web e identificar estruturas que possam ser caracterizadas como (a) Entidades Nomeadas (ENs) e (b) Padrões Textuais (PTs). Para que o aprendizado sem fim seja viável é fundamental que sejam implantadas estratégias que permitam que o

conhecimento adquirido possa ser reutilizado, dando continuidade e sequenciamento ao aprendizado.

Presentemente o RTWP opera como descrito a seguir. Além da ontologia inicial, é fornecido ao sistema um conjunto de entradas (o valor *default* é de 10 ENs e 10 PTs (com 1 argumento), para cada categoria participante da ontologia), que será utilizado pelo sistema na identificação de novas ENs e novos PTs em textos disponíveis na Web em português. Suponha, por exemplo, que *cidade* seja uma categoria da ontologia. Sempre (ou quase sempre) que o sistema encontra a sentença *"X é uma cidade localizada..."* (que caracteriza um dos 10 PTs da categoria *cidade*) a cadeia *X* refere-se a uma cidade. Assim, o sistema passa ler a Web em busca deste PT, para extrair a cadeia *X*. Suponha que após algum processamento as cadeias *"São Paulo", "São Carlos"* e *"Curitiba"* tenham sido encontradas; tais cadeias são nomeadas ENs candidatas (à promoção como legítimas). As ENs promovidas são aquelas que possuem maior acurácia e, consequentemente, são aceitas como corretas pelo sistema. As ENs promovidas são usadas no aprendizado de novos PTs e estes, usados no aprendizado de novas ENs, dando continuidade às iterações do sistema. No trabalho descrito neste artigo foi usado o *Yahoo Boss* para a recuperação de páginas Web. O *Yahoo Boss* é uma API gratuita, acessível a qualquer usuário via o *site* http://developer.yahoo.com/search/boss/.

Os três tipos de acoplamento previstos no trabalho são informalmente apresentados e discutidos a seguir. Acoplamentos caracterizados como Tipo1 são a forma mais simples de acoplamento e são realizados usando uma mesma categoria e não necessariamente o mesmo padrão textual (PT). Por meio deles, a partir de ENs são extraídos PTs e vice-versa; esse processo envolve uma tarefa de classificação. Como exemplo pode-se ter: X_1 = "São Carlos" e X_2 = "é uma cidade", em que X_1 é classificado como cidade e X_2 é classificado como um padrão textual (PT) de cidade. Na eventualidade de X1 ser classificado como não sendo cidade e X2 como um PT de cidade, ambos são penalizados por meio de alteração dos respectivos *scores*.

Acoplamentos caracterizados como do Tipo2 fazem uso dos chamados exemplos negativos (categorias que são mutuamente exclusivas com relação à categoria corrente). O Tipo2 acontece no domínio das ENs e no domínio dos PTs separadamente e, a partir dai, segue o mesmo processo implementado pelo Tipo1. Por exemplo, considere X_1 = "São Carlos", que será entrada tanto para classificadores de ENs associados à categoria *cidade* quanto para os classificadores de ENs associados às categorias mutuamente exclusivas com *cidade* (por exemplo, a categoria *país* e a categoria *continente*). Caso X_1 seja classificado como *cidade* e não seja classificado como *país* ou como *continente*, X_1 tem o seu *score* (associado à categoria *cidade*) aumentado, caso contrário, diminuído. Processo similar é realizado com PTs.

Acoplamentos que se caracterizam como do Tipo3 estão associados ao aprendizado de relações semânticas (RS), que são parte da ontologia inicial. Por *default*, o conjunto de entrada associado a cada RS é constituído por 10 pares de ENs e 10 PTs de dois argumentos. O acoplamento Tipo3 agrega o Tipo1 e o Tipo2 ao aprendizado de relações semânticas. Antes da realização da classificação da relação semântica, cada um dos dois argumentos da relação é avaliado pelos classificadores associados à categoria à qual o argumento em questão pertence. Dependendo do *score* dessas avaliações, é ativado o processo de classificação da relação semântica.

6. Uso de Acoplamentos: Experimentos e Discussão dos Resultados

Considerando a ontologia adotada (inspirada na descrita em [Betteridge et al. 2009b]), os experimentos descritos a seguir foram realizados para exame do impacto do uso dos acoplamentos propostos, na minimização do problema de desvio de conceito. As categorias consideradas nos experimentos são: cidade, companhia, setor econômico, pessoa e equipe esportiva (i.e., as participantes da ontologia inicial fornecida ao sistema). É importante ressaltar que nem todas as categorias que se encontram descritas na referência acima, foram contempladas no presente estudo.

Os resultados foram obtidos por meio da comparação das taxas de acerto obtidas com o uso de acoplamento *versus* sem o uso de acoplamento. Todos os experimentos descritos implementam o uso de acúmulo de conhecimento. Acumular conhecimento significa que todo o conhecimento (candidatos e promovidos) em uma iteração, será transferido às próximas. A precisão e cobertura foram calculadas de forma simplificada, considerando apenas o conhecimento extraído e não todo o conhecimento disponível. O valor da precisão foi calculado como a razão entre o número de promoções corretas e o número total de promoções. Já a cobertura representa apenas o número total de promoções.

Foram realizados experimentos com diferentes parâmetros de promoção; no caso isso se refere ao número máximo permitido de PTs (NMPTs) que podem ser promovidos (escolhido empiricamente como 10 e 3). Para a promoção de ENs, uma lista ordenada pelo valor do *score* é produzida e as ENs que estão no 1/3 do topo da lista são promovidas (note que o tamanho máximo dessa lista, presentemente, é de 500 elementos). Os resultados mostrados são referentes à execução do RTWP por 5 iterações.

Para os experimentos com NMPTs=10 os resultados mostraram tendência à melhoria de confiabilidade no conhecimento promovido com o uso dos acoplamentos; porém devido ao grande esforço computacional exigido, os experimentos com esse valor foram executados apenas até a quarta iteração, como mostra a Figura 1(a) e se limitam à essa figura. Esses resultados são exibidos apenas com o intuito de evidenciar o quanto a propagação de erros, causada por desvio de conceito, sem o uso de acoplamento, pode afetar o desempenho do sistema.

Os experimentos com NMPTs=3 exibem ainda a tendência na melhoria da precisão com o uso dos acoplamentos, como mostra a Figura 1(b). Contrário à situação anterior, o sistema foi executado até o número máximo de iterações planejado. Nas duas figuras são exibidos os resultados sem e com acoplamentos (Tipo1 e Tipo2) levando em consideração a quantidade de erros e de acertos, em cada experimento. Pode ser notado que o uso do acoplamento, em ambos os casos, melhorou a precisão e a cobertura. Com relação à categoria *pessoa*, por exemplo, como pode ser visto na Figura 2, os resultados obtidos mostram a contribuição do uso de acoplamentos.

Tabela 1 e Tabela 2 apresentam o resultado individualizado por iteração e por categoria, para os experimentos com e sem acoplamentos, respectivamente. Tais resultados evidenciam que uma definição consistente de ontologia aliada ao uso de acoplamentos promovem melhoria dos resultados quanto à cobertura e precisão.

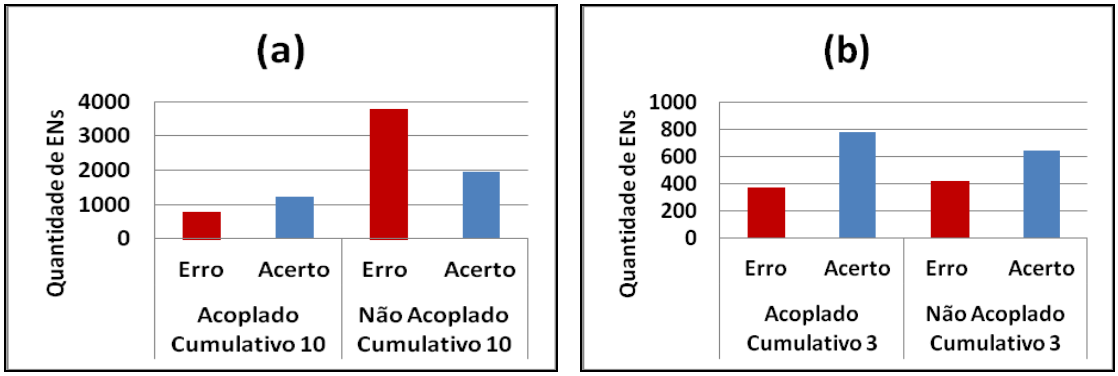

Figura 1. (a) Aprendizado Acoplado e Não Acoplado Cumulativo (NMPTs=10) e (b) Aprendizado Acoplado e Não Acoplado Cumulativo (NMPTs=3)

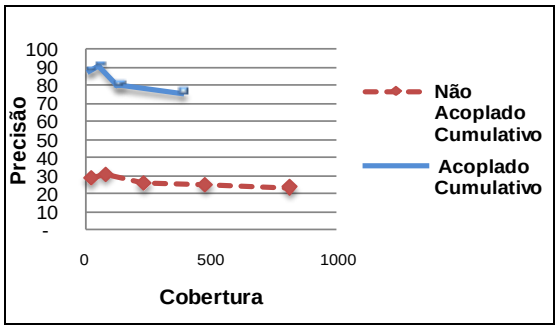

Figura 2. Aprendizado Acoplado e Não Acoplado Cumulativo relativos à categoria *pessoa*

Analisando os resultados nas tabelas 1-2, as categorias tiveram melhor desempenho quanto à precisão e cobertura com o uso do acoplamentos (Tipo1 e Tipo2). Note em ambas as tabelas algumas particularidades: (1) para as categorias *cidade*, *companhia* e *equipe esportiva*, o desempenho com relação à precisão e cobertura tende a ser melhor com o uso de acoplamentos. Para a categoria *setor econômico* o uso de acoplamentos permitiu o aprendizado de apenas 9 instâncias e, sem o uso de acoplamentos, 49. Entretanto, sem o uso de acoplamentos, a precisão caiu para 61% *versus* 89% obtida com o uso de acoplamentos. Para a categoria *pessoa*, o comportamento em números é similar ao da categoria *setor econômico*. Esses valores caracterizam um aspecto bastante importante de sistemas de aprendizado sem fim: o da promoção da corretude do conhecimento adquirido em detrimento à promoção do volume de conhecimento adquirido.

O acoplamento Tipo3 obteve bons resultados em todas as categorias, porém a cobertura foi muito baixa como mostra a Figura 3.

Tabela 1. Resultados por Categoria com Acoplamento Tipo2

Iteração		1	2	3	4	5
cidade	Qt Acerto	33	69	99	126	169
	%Acerto	94	93	95	94	93
companhia	Qt Acerto	14	24	34	42	60
	%Acerto	100	88	82	81	85
setor econômico	Qt Acerto	4	9	9	9	9
	%Acerto	75	89	89	89	89
pessoa	Qt Acerto	17	58	126	137	377
	%Acerto	88	91	80	80	76
equipe esportiva	Qt Acerto	14	25	32	39	62
	%Acerto	100	84	88	85	76

Tabela 2. Resultados por Categoria sem Acoplamento

Iteração		1	2	3	4	5
cidade	Qt Acerto	32	61	88	109	128
	%Acerto	91	92	92	92	93
companhia	Qt Acerto	10	20	27	33	45
	%Acerto	90	80	78	73	67
setor econômico	Qt Acerto	5	11	19	35	49
	%Acerto	100	100	74	71	61
pessoa	Qt Acerto	21	79	231	475	810
	%Acerto	28	31	25	25	23
equipe esportiva	Qt Acerto	14	23	30	36	48
	%Acerto	100	100	97	86	73

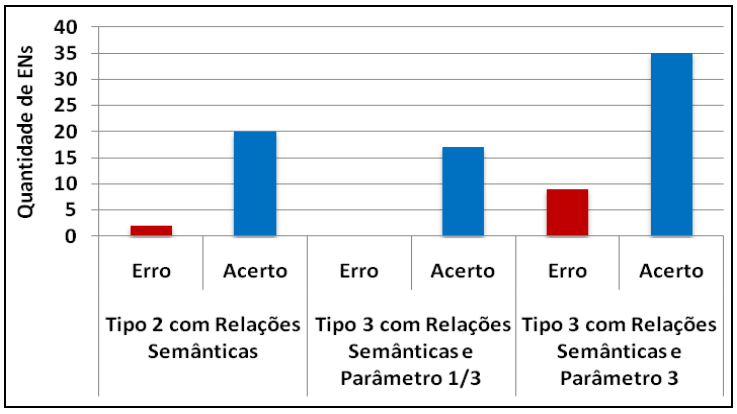

Figura 3. Aprendizado com Acoplamento Tipo2 e Acoplamento Tipo3

7. Conclusão

Os resultados obtidos evidenciam que o uso de diversos tipos de acoplamento tende a promover a precisão e a cobertura do conhecimento aprendido. Isso mostra que o sistema melhora seu desempenho quanto aprende várias categorias ao mesmo tempo. O uso de vários acoplamentos mostrou a tendência, também, do sistema manter por mais tempo o valor das precisões enquanto que, sem o uso, as quedas foram bruscas. Com a minimização do desvio de conceito o desempenho do sistema melhora. Um desvio de conceito alto é indicativo que o sistema não conseguiu aprender corretamente após muitas iterações; essa situação ocorreu com algumas categorias, como descrito na Seção 6. Não pode ser esquecido que a organização da ontologia tem papel crucial no desempenho do sistema; uma ontologia pobremente estruturada vai promover o aumento do desvio de conceito, mesmo que acoplamentos tenham sido usados.

Futuramente o acoplamento Tipo3 será refinado com vistas à promoção de precisão e cobertura e o acoplamento Tipo4 (agregado de acoplamentos) será implementado.

Agradecimentos

Os autores agradecem à Yahoo e às agências de fomento CNPq e CAPES pelo apoio fornecido.

Referências

Banko, M.; Cafarella, M. J.; Soderland, S.; Broadhead, M.; Etzioni, O. (2007) "Open information extraction from the Web", In: International Joint Conference on Artificial Intelligence, California: Morgan Kaufmann Publishers Inc., p. 2670-2676.

Banko, M. and Etzioni, O. (2008) "The tradeoffs between open and traditional relation extraction", In: Annual Meeting of the Association for Computational Linguistics, Philadelphia: Association for Computational Linguistics, p. 28-36.

Betteridge, J.; Carlson, A.; Hong, S. A. ; Hruschka Jr., E. R.; Law, E. L. M.; Mitchell, T.; Wang, S. (2009a) "Toward never ending language learning", In: AAAI 2009 Spring Symposium on Learning by Reading and Learning to Read, Palo Alto: Association for the Advancement of Artificial Intelligence, p. 1-2.

Betteridge, J.; Carlson, A.; Hruschka Jr., E. R.; Mitchell, T. M. (2009b) "Coupling semi-supervised learning of categories and relations", In: The NAACL HLT Workshop on Semi-supervised Learning for Natural Language Processing, Association for Computational Linguistics, p. 1-9.

Blum, A. and Mitchell, T. (1998) "Combining labeled and unlabeled data with co-training", In: Proceedings of the Annual Conference on Computational Learning Theory (COLT), Madison: ACM, p. 92-100.

Curran, J. R.; Murphy, T.; Scholz, B. (2007) "Minimising semantic drift with mutual exclusion bootstrapping", In: Proceedings of Pacific Association for Computational Linguistics Conference, Melbourne, Australia, p. 172–180.

Carlson, A. (2010) Coupled Semi-Supervised Learning, PhD. Thesis – School of Computer Science, Carnegie Mellon University, Pittsburgh, USA.

Carlson, A.; Betteridge, J.; Kisiel, B.; Settles, B.; Hruschka Jr., E. R.; Mitchell, T. M. (2010) "Toward an architecture for never-ending language learning", In: Proceedings of the Conference on Artificial Intelligence (AAAI), p. 1306-1313.

Chapelle, O.; Schölkopf, B.; Zien, A. (2006) Semi-Supervised Learning, Cambridge, MA: MIT Press, 2006.

Etzioni, O.; Banko, M.; Cafarella, M. J. (2006) "Machine reading", In: Proceedings of The 21st. National Conference on Artificial Intelligence (AAAI), p. 1517-1519.

Etzioni, O.; Fader, A.; Christensen, J.; Soderland, S.; Mausam (2011) "Open information extraction: the second generation", In: Proceedings of the International Joint Conference on Artificial Intelligence, p. 3-10.

Mitchell, T. M.; Betteridge, J.; Carlson, A.; Hruschka, E.; Wang, R. (2009) "Populating the semantic web by macro-reading internet text", In: International Semantic Web Conference, Chantilly: Springer-Verlag, p. 998-1002.

Suchanek, F.M.; Kasneci, G.; Weikum, G. (2007) "Yago: a core of semantic knowledge", In: Proceedings of The 16th. International Conference on World Wide Web, ACM, New York, USA, p. 697-706.

Whitelaw, C.; Kehlenbeck, A.; Petrovic, N.; Ungar, L. (2008) "Web-scale named entity recognition In: Proceeding of the 17th ACM conference on Information and knowledge management (CIKM '08), New York, NY, USA, 123-132.

Yarowsky, D. (1995) "Unsupervised word sense disambiguation rivaling supervised methods", In: Proceedings of the Annual Meeting on Association for Computational Linguistics, Cambridge: Association for Computational Linguistics, p. 189-196.

Zhu, X.; Goldberg, A. B.; Brachman, R.; Dietterich, T. (2009) Introduction to semi-supervised learning, San Rafael, California, USA: Morgan and Claypool Publishers, 130 p.