Dimensionando para corpora muito, muito grande para Desambiguação da Linguagem Natural

Michele Banko e Eric Brill
Pesquisa da Microsoft
1 Maneira da Microsoft
Redmond, WA 98052 EUA
{mbanko,brill}@microsoft.com

Abstrato

A quantidade de texto on-line prontamente disponível atingiu centenas de bilhões de palavras e continua a aumentar.

crescer. No entanto, para a maioria dos núcleos naturais tarefas de linguagem, os algoritmos continuam a ser otimizados, testados e comparados após o treinamento em corpora consistindo de apenas um milhão de palavras ou menos. Neste artigo, avaliamos o desempenho de diferentes

métodos em um protótipo natural
tarefa de desambiguação de linguagem,
confusão define desambiguação, quando treinado em ordens de magnitude mais
dados rotulados do que os usados anteriormente. Temos a sorte de que, para esta aplicação

específica, corretamente
os dados de treinamento rotulados são gratuitos. Desde isso muitas vezes não será o caso, examinamos métodos para efetivamente
explorar corpora muito grandes quando dados rotulados têm um custo.

1 Introdução

Técnicas de aprendizado de máquina, que aprender automaticamente informações linguísticas de corpora de texto on-line, foram aplicados a uma série de problemas de linguagem natural ao longo da última década. Um grande percentual de artigos publicados nesta área envolvem comparações de diferentes aprendizagens abordagens treinadas e testadas com corpora comumente usados. Embora a quantidade de texto online disponível tenha aumentado a um ritmo dramático, o tamanho dos corpora de formação normalmente utilizados para a aprendizagem não aumentou. Em parte, isso se deve ao padronização de conjuntos de dados usados na área, bem como o custo potencialmente grande de anotar dados para os métodos de aprendizagem que dependem de texto rotulado.

A comunidade empírica da PNL colocou um esforço substancial na avaliação do desempenho de um grande número de aprendizado de máquina métodos sobre conjuntos de dados fixos e relativamente pequenos. No entanto, como agora temos acesso a significativamente mais dados, é de se perguntar que conclusões tiradas em pequenos conjuntos de dados podem ser mantidas quando estes métodos de aprendizagem são treinados usando muito corpora maiores.

Neste artigo, apresentamos um estudo dos efeitos do tamanho dos dados no aprendizado de máquina para desambiguação de linguagem natural. Em particular, estudamos o problema da seleção entre palavras confusas, usando ordens de magnitude mais dados de treinamento do que jamais foram aplicados a este problema. Primeiro mostramos o aprendizado curvas para quatro aprendizados de máquina diferentes algoritmos. Além disso, consideramos a eficácia da votação, seleção da amostra e aprendizagem parcialmente não supervisionada com grandes corpora de treinamento, na esperança de poder obter os benefícios que advêm corpora de formação significativamente maiores sem incorrer em custos demasiado elevados.

2 Desambiguação do conjunto de confusão

A desambiguação do conjunto de confusão é o problema de escolher o uso correto de uma palavra, dado um conjunto de palavras com o qual é comumente confuso. Exemplos de conjuntos de confusão incluem: {principle, principal}, {then, than}, {to,two,too} e {weather,whether}.

Inúmeros métodos foram apresentado para desambiguação confusa. O conjunto mais recente de técnicas inclui

algoritmos multiplicativos de atualização de peso (Golding e Roth, 1998), semântica latente análise (Jones e Martin, 1997), aprendizagem baseada na transformação (Mangu e Brill, 1997), gramáticas diferenciais (Powers, 1997), listas de decisão (Yarowsky, 1994) e uma variedade de classificadores bayesianos (Gale et al., 1993, Golding, 1995, Golding e Schabes, 1996). Em todas essas abordagens, o problema é formulado da seguinte forma: Dado um conjunto de confusão específico (por exemplo, {to,two,too}), todos ocorrências de membros do conjunto de confusão no conjunto de teste são substituídas por um marcador; onde quer que o sistema veja esse marcador, ele deve decidir qual membro do conjunto de confusão escolher.

A desambiguação do conjunto de confusão faz parte de uma classe de problemas de linguagem natural envolvendo a desambiguação de um conjunto relativamente pequeno de alternativas com base no contexto da string em que o local de ambiguidade aparece.
Outros problemas incluem o sentido das palavras desambiguação, marcação de classes gramaticais e algumas formulações de fragmentação frasal. Um aspecto vantajoso do conjunto de confusão
A desambiguação, que nos permite estudar os efeitos de grandes conjuntos de dados no desempenho, é que os dados de treinamento rotulados são essencialmente gratuitos, uma vez que a resposta correta é aparente em qualquer coleção de texto razoavelmente bem editado.

3 Experimentos de curva de aprendizagem

Este trabalho foi parcialmente motivado pelo desejo de desenvolver uma gramática melhorada. verificador. Dado um período fixo de tempo, consideramos qual seria a maneira mais eficaz de concentrar nossos esforços para obter a maior melhoria de desempenho. Alguns possibilidades incluíam modificar o padrão algoritmos de aprendizagem, explorando novas aprendizagens técnicas e usando técnicas mais sofisticadas características. Antes de explorar estes um pouco caminhos caros, decidimos primeiro ver o que aconteceria se simplesmente treinássemos um existente método com muito mais dados. Isto levou à exploração de curvas de aprendizagem para vários algoritmos de aprendizado de máquina: winnow1 , perceptron, Bayes ingênuo e um aluno muito simples baseado na memória. Para os três primeiros alunos, usamos a coleção padrão de recursos empregados para este problema: o conjunto de palavras dentro de uma janela da palavra alvo e colocações contendo palavras e/ ou classes gramaticais. O aluno baseado em memória usado apenas a palavra antes e a palavra depois como características.

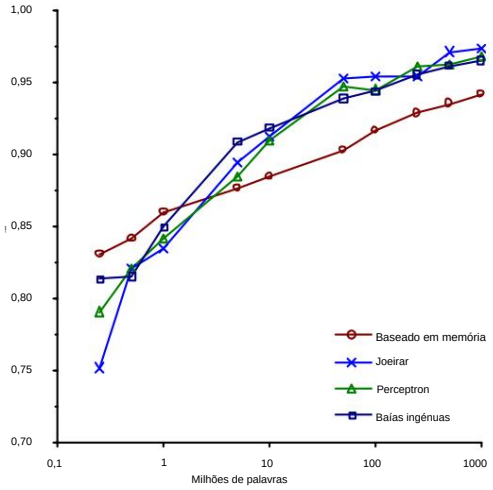

Figura 1. Curvas de aprendizagem para desambiguação de conjuntos de confusão

Coletamos um corpus de treinamento de 1 bilhão de palavras de uma variedade de textos em inglês, incluindo artigos de notícias, resumos científicos, transcrições governamentais, literatura e outros variadas formas de prosa. Este corpus de treinamento é três ordens de magnitude maior que o maior corpus de treinamento usado anteriormente para este problema. Usamos 1 milhão de palavras do texto do Wall Street Journal como conjunto de teste, e nenhum dado do Wall Street Journal foi usado na construção do corpus de treinamento. Cada aluno foi treinado em vários pontos de corte no corpus de treinamento, ou seja, o primeiro milhão palavras, os primeiros cinco milhões de palavras e assim por diante, até que um bilhão de palavras fossem usadas para treinamento. Para evitar vieses de treinamento que podem resultar da mera concatenação dos diferentes fontes de dados para formar um corpus de treinamento maior, construímos cada treinamento do corpus por meio de amostragem probabilística de sentenças de diferentes fontes ponderadas pelo tamanho de cada fonte.

Na Figura 1, mostramos curvas de aprendizagem para cada aluno, até um bilhão de palavras de

[1] Obrigado a Dan Roth por disponibilizar o Winnow e o Perceptron.

dados de treinamento. Cada ponto no gráfico é o desempenho médio em dez conjuntos de confusão para esse tamanho de corpus de treinamento. Observe que as curvas parecem ser log-lineares mesmo com um bilhão de palavras.

Claro que para muitos problemas, dados de treinamento adicionais têm um custo diferente de zero. No entanto, estes resultados sugerem que podemos querer reconsiderar o compromisso entre gastando tempo e dinheiro em algoritmo desenvolvimento versus gastá-lo em corpus desenvolvimento. Pelo menos para o problema da desambiguação confusa, nenhum dos alunos testados está próximo da assíntota em desempenho no tamanho do corpus de treinamento comumente empregado pela área.

Tais ganhos em precisão, no entanto, não são gratuitos. A Figura 2 mostra que o tamanho das representações aprendidas cresce linearmente em função do tamanho dos dados de treinamento. Para algumas aplicações, isso não é necessariamente uma preocupação. Mas para outros, onde o espaço é precioso, a obtenção dos ganhos que advêm de mil milhões de palavras de dados de treino pode não ser viável sem um esforço feito para comprimir a informação. Nesses casos, pode-se considerar vários métodos para compactar dados (por exemplo, Dagan e Engleson, 1995, Weng, et al, 1998).

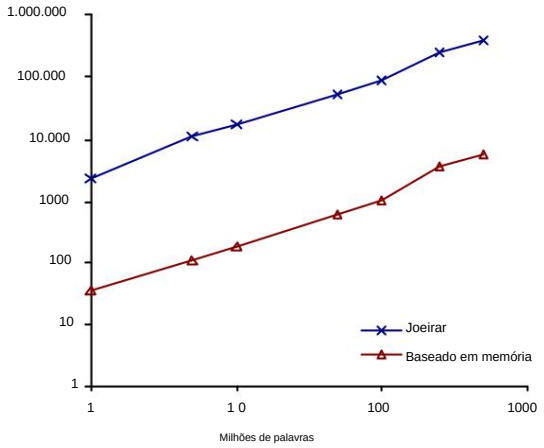

Figura 2. Tamanho da representação versus tamanho do corpus de treinamento

4 A eficácia da votação

A votação provou ser uma técnica eficaz para melhorar a precisão do classificador para muitos aplicações, incluindo marcação de classe gramatical (van Halteren, et al, 1998), análise (Henderson e Brill, 1999) e desambiguação do sentido das palavras (Pederson, 2000). Treinando um conjunto de classificadores em um único corpus de treinamento e depois combinando seus resultados em classificação, muitas vezes é possível atingir a precisão desejada com menos dados de treinamento rotulados do que seria necessário se apenas um classificador estivesse sendo usado. A votação pode ser eficaz na redução tanto do preconceito de um corpus de formação específico como do preconceito de um aluno específico. Quando um corpus de formação é muito pequeno, há muito mais espaço para que estes preconceitos venham à tona e, portanto, para que a votação seja eficaz. Mas será que a votação ainda oferece ganhos de desempenho quando os classificadores são treinados em corpora muito maiores?

A complementaridade entre dois alunos foi definido por Brill e Wu (1998) para quantificar a porcentagem de tempo em que um sistema está errado, que outro o sistema está correto e, portanto, fornece um limite superior na precisão da combinação. Como o tamanho do treinamento aumenta significativamente, esperaríamos que a complementaridade entre os classificadores diminuísse. Isto se deve em parte ao fato de que um corpus de treinamento maior reduzirá a variância do conjunto de dados e qualquer viés daí resultante. Além disso, algumas das diferenças entre os classificadores podem ser devidas à forma como eles lidam com um ambiente es conjunto de treinamento. Como resultado da comparação de uma amostra de dois alunos em função de conjuntos de formação cada vez maiores, vemos na Tabela 1 que a complementaridade realmente diminuir à medida que o tamanho do treinamento aumenta.

Tamanho do treinamento (palavras)	Complementaridade (L1,L2)
106	0,2612
107	0,2410
108	0,1759
10^9	0,1612

Tabela 1. Complementaridade

Em seguida, testámos se esta diminuição na complementaridade significava que a votação perdeva a sua eficácia à medida que o conjunto de formação aumentava. Para examinar o impacto da votação ao usar um corpus de treinamento significativamente maior, analisamos 3 dos 4 alunos em nosso conjunto de 10 pares confundíveis, excluindo os baseados em memória

aluno. A votação foi feita combinando a pontuação normalizada que cada aluno atribuiu a uma escolha de classificação. Na Figura 3, mostramos a precisão obtida na votação, juntamente com
a melhor precisão do aluno em cada
tamanho do conjunto de treinamento. Vemos que para corpora muito pequenos, votar é benéfico, resultando em melhor desempenho do que qualquer classificador único. Além de 1 milhão de palavras, pouco se ganha votando e, de fato, nos maiores conjuntos de treinamento, a votação prejudica a precisão.

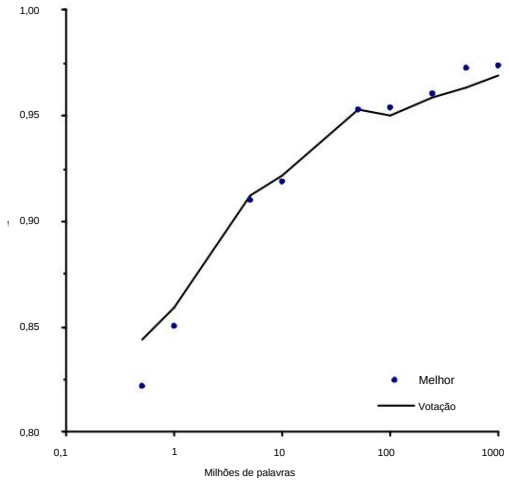

Figura 3. Votação entre classificadores

5 Quando os dados anotados não são Livre

Embora a observação de que as curvas de aprendizagem não são assíntotas, mesmo com ordens de
Embora a magnitude de mais dados de treinamento do que os usados atualmente seja muito interessante, esse resultado pode ter ramificações um tanto limitadas. Existem muito poucos problemas para os quais dados anotados deste tamanho estejam disponíveis gratuitamente. Certamente nós não podemos razoavelmente esperar que o manual a anotação de um bilhão de palavras junto com as árvores de análise correspondentes ocorrerá em breve (mas veja (Banko e Brill 2001) para uma discussão de que isso pode não ser completamente inviável). Apesar desta armadilha, existem
técnicas que podemos usar para tentar obter os benefícios de um treinamento consideravelmente maior corpora sem incorrer em
custos adicionais. Nas seções a seguir, estudamos duas dessas soluções: aprendizagem ativa e aprendizagem não supervisionada.

5.1 Aprendizagem Ativa

O aprendizado ativo envolve a seleção inteligente de uma porção de amostras para anotação de um conjunto de amostras de treinamento ainda não anotadas.
Nem todas as amostras em um conjunto de treinamento são igualmente úteis. Ao concentrar os esforços de anotação humana nas amostras de maior utilidade para o algoritmo de aprendizado de máquina, pode ser possível obter um melhor desempenho para um custo fixo de anotação do que se as amostras fossem escolhido aleatoriamente para anotação humana.

A maioria das abordagens de aprendizagem ativa funciona primeiro treinando um aluno inicial (ou família de alunos) e depois executando o(s) aluno(s) em um conjunto de amostras não rotuladas. Presume-se que uma amostra seja mais útil para treinamento quanto mais incerto for seu rótulo de classificação.

A incerteza pode ser julgada pelo relativo pesos atribuídos a diferentes rótulos por um único classificador (Lewis e Gale, 1994). Outra abordagem, amostragem baseada em comitê, primeiro cria um comitê de classificadores e então julga a incerteza da classificação de acordo com o quanto os alunos diferem entre os rótulos
atribuições. Por exemplo, Dagan e Engelson (1995) descreve uma técnica de amostragem baseada em comitê onde uma classe gramatical tagger é treinado usando um corpus de sementes anotado. Uma família de taggers é então gerada permutando aleatoriamente o tagger
probabilidades, e a disparidade entre as tags produzidas pelos membros do comitê é usada como uma medida de incerteza de classificação.
Sentenças para anotação humana são desenhadas, tendenciosamente a preferir aquelas que contêm alto instâncias de incerteza.

Embora tenha sido demonstrado que a aprendizagem ativa funciona para uma série de tarefas, a maioria das experimentos de aprendizagem ativa em natural o processamento de linguagem foi conduzido usando corpora sementes muito pequenos e conjuntos de exemplos não rotulados. Portanto, desejamos explorar situações em que temos ou podemos
pagar, um corpus de treinamento de tamanho não negligenciável (como para marcação de classes gramaticais) e ter acesso a grandes quantidades de dados não rotulados.

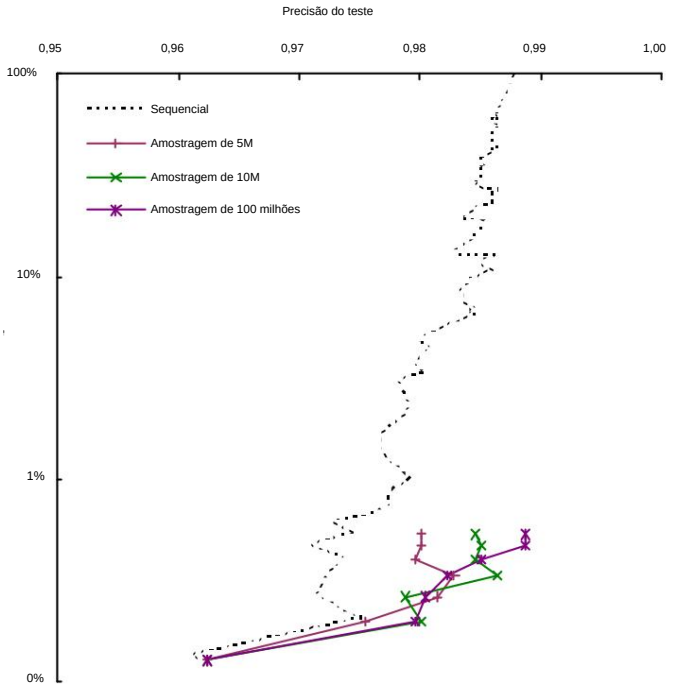

Figura 4. Aprendizagem Ativa com Grandes Corpora

Podemos usar bagging (Breiman, 1996), uma técnica para gerar um comitê de classificadores, para avaliar a incerteza do rótulo de uma potencial instância de treinamento. Com o ensacamento, uma variante do conjunto de treinamento original é construído por amostragem aleatória de sentenças com substituição do conjunto de treinamento de origem, a fim de produzir N novos conjuntos de treinamento de tamanho igual ao original. Após os N modelos terem sido treinados e executados no mesmo conjunto de testes, suas classificações para cada sentença de teste podem ser comparadas quanto à concordância de classificação. Quanto maior a discordância entre os classificadores, mais útil seria ter uma instância rotulada manualmente.

Usamos o classificador Bayes ingênuo, criando 10 classificadores, cada um treinado em bolsas geradas a partir de um milhão de palavras iniciais de dados de treinamento rotulados. Apresentamos o algoritmo de aprendizagem ativa que usamos abaixo.

Inicializar: os dados de treinamento consistem em X palavras rotulado corretamente
Iterar:

1) Gere um comitê de classificadores usando bagging no conjunto de treinamento
2) Execute o comitê na parte não rotulada do conjunto de treinamento
3) Escolha M instâncias do conjunto não rotulado para rotulagem - escolha o M/2 com a maior entropia de votos e depois escolha outro M/2 aleatoriamente - e adicione ao conjunto de treinamento

---

Inicialmente tentamos selecionar os M exemplos mais incertos, mas isso resultou em uma amostra muito tendenciosa para o difícil instâncias. Em vez disso, escolhemos aleatoriamente metade de nossas amostras para anotação e a outra metade daquelas cujos rótulos estamos mais interessados. incerto, a julgar pela entropia dos votos atribuídos à instância pelo comitê. Isso está, na verdade, distorcendo nossa amostra em direção a instâncias em que os classificadores são mais incerto de.

Mostramos os resultados da amostra seleção para confusão definir desambiguação em Figura 4. A linha rotulada como "sequencial" mostra a precisão do conjunto de testes alcançada para diferentes porcentagens do conjunto de treinamento de um bilhão de palavras, onde as instâncias de treinamento são obtidas aleatoriamente. Realizamos três atividades de aprendizagem ativa

experimentos, aumentando o tamanho do corpus de treinamento total não rotulado do qual podemos escolher amostras a serem anotadas. Nos três casos, a seleção da amostra supera a amostragem sequencial. No final de cada execução de treinamento no gráfico, o mesmo número de amostras foi anotado para treinamento. No entanto, vemos que quanto maior o conjunto de candidatos

instâncias para anotação, melhor será o precisão resultante. Ao aumentar o conjunto de instâncias de treinamento não rotuladas para ativos aprendendo, podemos melhorar a precisão com apenas um custo fixo de anotação adicional. Assim, é possível beneficiar-se da disponibilidade de corpora extremamente grandes sem incorrer nos custos totais de anotação, tempo de treinamento e tamanho de representação.

5.2 Aprendizagem Fracamente Supervisionada

Embora a secção anterior mostre que podemos beneficiar de uma formação substancialmente maior corpora sem a necessidade de anotações manuais adicionais significativas, seria ideal se poderia melhorar a precisão da classificação usando apenas nosso corpus anotado de origem e o grande corpus não rotulado, sem exigir qualquer rotulagem manual adicional. Nesta seção nos voltamos para a aprendizagem não supervisionada na tentativa de atingir esse objetivo. Numerosas abordagens foram exploradas para explorar situações onde alguma quantidade de dados anotados está disponível e existe uma quantidade muito maior de dados sem anotação, por exemplo, treinamento de tagger de classe gramatical HMM de Marialdo (1994), Charniak experimento de reciclagem do analisador (1996), As sementes de Yarowsky para desambiguação do sentido das palavras (1995) e Nigam et al's (1998) classificador de tópicos aprendido em parte com documentos não rotulados. Uma boa discussão deste problema geral pode ser encontrada em Mitchell (1999).

A questão que queremos responder é se há algo a ganhar combinando atividades não supervisionadas e supervisionadas. aprendendo quando ampliamos tanto a semente corpus e o corpus não rotulado significativamente. Podemos novamente recorrer a um comité de ensacados classificadores, desta vez para aprendizagem não supervisionada. Considerando que com a aprendizagem ativa queremos escolher as instâncias mais incertas para anotação humana, com aprendizagem não supervisionada, deseja escolher as instâncias que possuem o maior probabilidade de estar correto para rotulagem automática e inclusão em nosso dados de treinamento rotulados.

Na Tabela 2, mostramos o conjunto de teste precisão (média dos quatro pares de confusão que ocorrem frequentemente) como um função do número de classificadores que concordam com o rótulo de uma instância. Por esta experimento, treinamos uma coleção de 10 classificadores Bayes ingênuos, usando ensacamento em um 1-corpus semente de um milhão de palavras. Como pode ser visto, quanto maior a concordância do classificador, mais é provável que uma amostra de teste tenha sido rotulada corretamente.

Classificadores De acordo	Teste Precisão
10	0,8734
9	0,6892
8	0,6286
7	0,6027
6	0,5497
5	0,5000

Tabela 2. Acordo do Comitê vs. Precisão

Como os casos em que todas as malas concordam têm a maior probabilidade de estarem corretos, tentamos aumentar automaticamente o nosso conjunto de treinamento rotulado usando o corpus inicial rotulado de 1 milhão de palavras junto com a coleção de classificadores Bayes ingênuos descritos acima. Todas as instâncias do restante do corpus com as quais todos os 10 classificadores concordaram foram selecionadas, confiando no rótulo acordado. Os classificadores foram então retreinados usando a semente rotulada corpus mais o novo material de treinamento coletado automaticamente na etapa anterior.

Na Tabela 3 mostramos os resultados de esses experimentos de aprendizagem não supervisionados para dois conjuntos de confusão. Em ambos os casos, ganhamos com o treinamento não supervisionado em comparação com o uso apenas do corpus de sementes, mas apenas até certo ponto. Neste ponto, a precisão do conjunto de testes começa a diminuir à medida que instâncias de treinamento adicionais são colhido automaticamente. Somos capazes de atingir melhorias na precisão gratuitamente usando aprendizado não supervisionado, mas ao contrário de nossos experimentos de curva de aprendizado usando rótulos corretos dados, a precisão não continua a melhorar com dados adicionais.

	{depois do que}		{entre, entre}	
	Teste	%Total	dados	% Total de
	Precisão	Dados de treinamento	Precisão	de treinamento de teste
106 -wd semente de corpo de semente	0,9624	0,1	0,8183	0,1
rotulada+5x106 wds, semente não	0,9588	0,6	0,8313	0,5
supervisionada+107 wds, semente não	0,9620	1,2	0,8335	1,0
supervisionada+108 wds, semente não	0,9715	12,2	0,8270	9.2
supervisionada+5x108 wds, não	0,9588	61.1	0,8248	42,9
supervisionado 109 wds, supervisionado	0,9878	100	0,9021	100

Tabela 3. Aprendizagem Não Supervisionada Baseada em Comitê

Charniak (1996) realizou um experimento no qual treinou um analisador em um milhão de palavras de dados analisados, executei o analisador em mais 30 milhões de palavras e usei o análises resultantes para reestimar o modelo probabilidades. Fazer isso deu um pequeno melhoria em relação apenas ao uso manual dados analisados. Repetimos esta experiência com nossos dados e mostre o resultado na Tabela 4. Escolher apenas as instâncias rotuladas com maior probabilidade de serem corretas, conforme julgado por um comitê de classificadores, resulta em maior precisão do que usar todas as instâncias classificadas por um modelo treinado com o corpus de sementes rotulado.

	Não supervisionado: Todos os rótulos	Não supervisionado: A maioria dos rótulos certos
	{depois do que}	
107 palavras	0,9524	0,9620
108 palavras	0,9588	0,9715
5x108 palavras	0,7604 0,9588	
	{entre, entre}	
107 palavras	0,8259 0,8335	
108 palavras	0,8259 0,8270	
5x108 palavras	0,5321	0,8248

Tabela 4. Comparação de métodos de aprendizagem não supervisionados

Ao aplicar a aprendizagem não supervisionada para melhorar um método treinado em sementes, constatou consistentemente uma melhoria desempenho seguido por um declínio. Isto provavelmente se deve ao fato de eventualmente ter chegado a um ponto em que os ganhos de dados de treinamento adicionais são compensados pelo viés da amostra na mineração dessas instâncias. Talvez seja possível combinar aprendizagem ativa com aprendizagem não supervisionada aprendizagem como forma de reduzir esse viés amostral e obter os benefícios de ambas as abordagens.

6 Conclusões

Neste artigo, analisamos o que acontece quando começamos a aproveitar as grandes quantidades de texto que agora estão prontamente disponíveis. Mostramos que por um classificação prototípica de linguagem natural tarefa, o desempenho dos alunos pode se beneficiar significativamente de conjuntos de treinamento muito maiores. Também mostramos que tanto a aprendizagem ativa quanto a aprendizagem não supervisionada podem ser usadas para obter pelo menos algumas das vantagens que vem com dados de treinamento adicionais, ao mesmo tempo que minimiza o custo de recursos humanos adicionais anotação. Propomos que um próximo passo lógico para a comunidade de pesquisa seria direcionar esforços para aumentar o tamanho das coleções de treinamento anotadas, ao mesmo tempo em que desvalorizava o foco na comparação diferentes técnicas de aprendizagem treinadas apenas em pequenos corpora de treinamento. Enquanto é encorajando o facto de existir uma grande quantidade de texto online, ainda há muito trabalho a ser feito se quisermos aprender a melhor forma de explorar este recurso para melhorar o processamento da linguagem natural.

Referências

Banko, M. e Brill, E. (2001). *Mitigando o Problema de escassez de dados.* Em Anais da Conferência sobre Tecnologia da Linguagem Humana, San Diego, Ca.

Breiman L., (1996). *Preditores de ensacamento,* aprendizado de máquina 24 123-140.

Brill, E. e Wu, J. (1998). *Combinação de classificadores para melhor desambiguação lexical.* Em Anais da 17ª Internacional Conferência sobre Lingüística Computacional.

Charniak, E. (1996). *Gramáticas de Treebank,* Procedimentos AAAI-96 , Menlo Park, Califórnia.

Dagan, I. e Engelson, S. (1995). *Amostragem baseada em comitê para treinamento probabilístico*

classificadores. Em Proc. ML-95, o 12º Int. Conf. em aprendizado de máquina.

Gale, WA, Church, KW e Yarowsky, D. (1993). *Um método para desambiguar palavras sentidos em um grande corpus.* Os computadores e o Humanidades, 26:415--439.

Golding, AR (1995). *Um método híbrido bayesiano para correção ortográfica sensível ao contexto.* Em Processo. 3º Workshop sobre Corpora Muito Grandes, Boston, MA.

Golding, AR e Roth, D. (1999), *Uma abordagem baseada em Winnow para correção ortográfica sensível ao contexto.* Aprendizado de Máquina, 34:107--130.

Golding, AR e Schabes, Y. (1996). *Combinação de métodos baseados em trigramas e recursos para correção ortográfica sensível ao contexto.* Em Proc. 34ª Reunião Anual da Associação de Linguística Computacional, Santa Cruz, CA.

Henderson, JC e Brill, E. (1999). *Explorando a diversidade no processamento de linguagem natural: combinando analisadores.* Em 1999, Sigdat Conjunto Conferência sobre Métodos Empíricos em Processamento de Linguagem Natural e Corpora Muito Grandes. ACL, New Brunswick, NJ. 187-194.

Jones, MP e Martin, JH (1997). *Correção ortográfica contextual usando semântica latente análise.* Em Proc. 5ª Conferência sobre Processamento de Linguagem Natural Aplicada, Washington, DC.

Lewis, DD e Gale, WA (1994). *Um algoritmo sequencial para treinamento* ***de classificadores de texto.*** Em Anais da 17ª Conferência Internacional Anual ACM/SIGIR, 3--11.

Mangu, L. e Brill, E. (1997). *Regra automática aquisição para correção ortográfica.* Em Proc. 14ª Conferência Internacional sobre Aprendizado de Máquina. Morgan Kaufmann.

Merialdo, B. (1994). *Marcação de texto em inglês com um modelo probabilístico.* Lingüística Computacional, 20(2):155--172.

Mitchell, TM (1999), *O papel dos dados não rotulados na aprendizagem supervisionada,* em Proceedings of the Sexto Colóquio Internacional sobre Cognição Ciência, San Sebastián, Espanha.

Nigam, N., McCallum, A., Thrun, S. e Mitchell, T. (1998). *Aprender a classificar texto de documentos rotulados e não rotulados.* Em Anais do Décimo Quinto Nacional Conferência sobre Inteligência Artificial. AAAI Imprensa..

Pedersen, T. (2000). *Uma abordagem simples para construir conjuntos de classificadores bayesianos ingênuos para desambiguação do sentido das palavras.* Em Anais do Primeira Reunião do Capítulo Norte-Americano da Association for Computational Linguistics, 1 a 3 de maio de 2000, Seattle, WA

Poderes, D. (1997). *Aprendizagem e aplicação de gramáticas diferenciais.* Em Proc. Reunião do Grupo de Interesse Especial da ACL em Natural Aprendizagem de línguas, Madrid.

van Halteren, H. Zavrel, J. e Daelemans, W. (1998). *Melhorando a classe de palavras baseada em dados marcação por combinação de sistema.* Em COLING ACL'98, páginas 491497, Montreal, Canadá.

Weng, F., Stolcke, A e Sankar, A (1998). *Representação e geração eficiente de rede.* Processo. Internacional Conf. na linguagem falada Processamento, vol. 6, páginas 2531-2534. Sidney, Austrália.

Yarowsky, D. (1994). *Listas de decisão para resolução de ambigüidades lexicais: aplicação ao acento restauração em espanhol e francês.* Em Proc. 32ª Reunião Anual da Associação de Lingüística Computacional, Las Cruces, NM.

Yarowsky, D. (1995) *Sentido de palavra não supervisionado desambiguação rivalizando com métodos supervisionados.* Nos Anais da 33ª Reunião Anual da Association for Computational Linguistics. Cambridge, MA, pp.