Análise e Implementação de Modelos Contextuais para Desambiguação de Entidades Nomeadas em Fluxos de Mensagens

Alexandre Davis[1], **Adriano C. M. Pereira**[1]

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)

{agdavis,adrianoc}@dcc.ufmg.br

Abstract. *Named entity disambiguation in message streams is a new challenge in natural language processing. The low informational rate and lack of syntactical structure in this kind of text may decrease the accuracy of traditional disambiguation approaches. In this paper, we propose to use contextual models of Twitter messages to minimize text usage. Our models are based on the behavior of social network users and on the instant in which the message has been posted. Our results show that these models perform better than approaches that consider only textual attributes.*

Resumo. *Desambiguar entidades em fluxos de mensagens extraídos de mídias sociais é um novo desafio na área de processamento de linguagem natural. O baixo teor informacional e a má formação sintática desse tipo de texto prejudicam a aplicação de abordagens clássicas para resolver este problema. Neste trabalho, é proposta uma modelagem contextual das mensagens de microblogs, de forma a aperfeiçoar os modelos tradicionais de desambiguação, que utilizam apenas texto. Os modelos são baseados no comportamento dos usuários na rede social e no instante em que a mensagem foi postada. Nossos resultados mostram que esses modelos superam abordagens que levam em consideração apenas atributos textuais.*

1. Introdução

A comunicação mediada pela internet (CMI) está sendo transformada pela ascensão das redes sociais, o que implica no surgimento de novos padrões de escrita. Nos microblogs, por exemplo, as mensagens têm tamanho restrito e podem ser afetadas por eventos reais e emoções pessoais. Dessa forma, enquanto os textos mais convencionais de IMC (e.g. e-mails, notícias) são, em geral, muito informativos, bem escritos e com um vocabulário constante, as mensagens de microblogs tendem a ser curtas, apresentar erros sintáticos e gramaticais e ser escritas com um vocabulário altamente dinâmico, influenciado por acontecimentos do mundo real. Essas características podem tornar impraticável a solução de problemas, como a desambiguação de entidades nomeadas (*named entity disambiguation*, NED), utilizando técnicas tradicionais de processamento de linguagem natural (*natural language processing*, NLP).

Neste trabalho, propomos novas técnicas para NED em tempo real em um fluxo de mensagens. Em nossa formulação, o fluxo de dados S é composto de mensagens que

contêm palavras de um conjunto W^i, que podem referenciar uma entidade e. Nosso objetivo é remover todas as mensagens em S que não referenciam e. Como pretendemos aplicar a técnica proposta em tempo real, consideramos não ser possível a consulta a dados externos (e.g. Wikipedia), devido restrições de tempo. É importante notar que o vocabulário dinâmico e o pequeno tamanho das mensagens diminuem a efetividade de modelos de linguagem tradicionais em NLP. Além disso, abordagens ingênuas de aprendizado de máquina falham nesse cenário, pois dados anotados são caros e o classificador precisa ser constantemente atualizado com novos exemplos.

Acreditamos que muitos desses problemas podem ser superados através do uso de informação contextual (i.e. não-textual), implícita no fluxo de mensagens. Com base na teoria da Pragmática [Yus 2011], também acreditamos que, na comunicação humana, existe uma tendência de deixar implícitas informações que os interlocutores, supostamente, seriam capazes de obter por conta própria . Por exemplo, quando duas pessoas estão conversando e uma diz: "Você viu a partida ontem?", quem pergunta espera que o interlocutor consiga inferir a qual partida ele se refere. Nesse caso, as informações contextuais implícitas podem ser o conhecimento das preferências esportivas do interlocutor e/ou das partidas que aconteceram no dia anterior. Generalizando, a localidade social (i.e. usuários têm características semelhantes às de seus interlocutoes) e a localidade temporal adicionam informações semânticas fundamentais para a tarefa de desambiguação no cenário de fluxo de mensagens.

Nossa hipótese principal é a de que quando usuários fazem referências ambíguas a uma entidade específica, eles estão supondo que seus interlocutores serão capazes de desambiguá-las utilizando informações contextuais. Para verificar essa hipótese, foram propostos e implementados métodos para extrair a localidade social e a localidade temporal de informações das mensagens do fluxo (e.g. menções, *timestamp*, usuário emissor) e do histórico de mensagens dos usuários. Com isso, foi possivel avaliar a efetividade dessa proposta para desambiguação de entidades.

Este trabalho foi integralmente idealizado e implementado pelo aluno Alexandre Davis, durante o seu último ano de graduação. Uma extensa pesquisa em Linguística e Pragmática foi realizada para obter as bases teóricas para este estudo. Durante 2012, foi publicado um artigo relacionado a este na conferência da *Association of Computational Linguistics* (Qualis A1) [Davis et al. 2012].

2. Trabalhos Relacionados

Desambiguação de entidades nomeadas é um assunto amplamente estudado na área de processamento de linguagem natural. As abordagens mais tradicionais realizam essa tarefa utilizando bases externas, como DBpedia [Bizer et al. 2009], YAGO [Suchanek et al. 2007] e, mais recentemente, AIDA [Yosef et al. 2011], para complementar as informações contidas no texto. Muitas dessas técnicas são aplicadas em textos jornalísticos [Cucerzan 2007, Hoffart et al. 2011], blogs e páginas Web [Wang et al. 2012].

É esperado que tais técnicas não tenham um bom desempenho em um cenário de textos mal estruturados, sintaticamente incorretos e de vocabulário altamente dinâmico como ocorre em textos extraídos de microblogs [Davis et al. 2012]. Além disso, argumentamos que não é possível realizar consultas a bases externas se desejamos realizar a tarefa de desambiguação em tempo real. Para lidar com tais problemas, foram propostas

técnicas para classificação textual de mensagens de microblog nas quais o conjunto de treinamento é gerado automaticamente [Davis et al. 2012].

Neste trabalho, são utilizados atributos não-textuais, implícitos no stream de mensagens. Esse tipo de atributo é fundamental para a comunicação humana, conforme mostram os trabalhos de Pragmática [Sperber and Wilson 1986], porém não vem sendo utilizado para obter o conteúdo semântico de mensagens extraídas de redes sociais. Alguns trabalhos já fizeram uso desse tipo de informação para determinar o viés de usuários em tópicos polêmicos como futebol e política[Guerra et al. 2011b, Guerra et al. 2011a]. Outra técnica que utiliza informações contextuais, no caso, localidade social, foi proposta por Nguyen et al. [2011] para melhorar a latência de sistemas de redes sociais.

3. Formulação do Problema

Neste trabalho, utilizamos fluxos de mensagens extraídos do Twitter (*Tweet streams*). Um tweet é modelado como uma tripla $t =< u, d, m >$, sendo u o usuário emissor, d o *timestamp* da messagem e m a mensagem textual. Para coletar o fluxo de tweets S, foi utilizada a função *filter* da API do Twitter. Dado um conjunto de palavras $W = \{w_1, w_2, ..., w_n\}$, esta função retorna o fluxo de tweets $S = \{t(u, d, m) \mid m \cap W \neq \phi\}$.

Nosso objetivo é extrair um fluxo S^e de mensagens relacionadas com uma entidade nomeada e. A forma mais trivial de realizar essa tarefa é selecionar um conjunto de palavras W^e e usar a função *filter* do Twitter. No entanto, escolher essas palavras não é simples. Se referências ambíguas a e não são cobertas em W^e, podemos encontrar baixas taxas de revocação. Por outro lado, incluir palavras ambíguas em W^e pode introduzir várias mensagens falsas-positivas em S^e e diminuir a precisão do fluxo extraído.

Uma das formas de lidar com esse compromisso é escolher W^e grande e filtrar todos os exemplos que não se referem à entidade e. Para isso, no entanto, é necessária uma técnica que seja capaz de desambiguar as referências ambíguas a e presentes em W^e. Analisamos a eficácia de atributos contextuais implícitos no fluxo de mensagens para realizar esse tipo de tarefa. Em seguida, será mostrada a base teórica utilizada para modelar esses atributos.

4. Modelo Proposto

O processo de comunicação humana é um fenômeno complexo e difícil de ser reproduzido computacionalmente. Buscando modelar esse fenômeno, a teoria de Pragmática propõe a divisão da comunicação em duas macro-etapas: decodificação e interpretação [Yus 2011]. Durante a primeira etapa, o receptor simplesmente reconhece o que foi dito, em outras palavras, é feito o reconhecimento de símbolos semânticos (palavras, sons, texto falado, etc.) e uma mensagem é gerada. Durante a segunda etapa, o interlocutor interpreta essa mensagem para descobrir o que o emissor quis dizer. Para isso, pode ser necessário utilizar informações contextuais, como conhecimento da personalidade e interesses do emissor, eventos que estão acontecendo no momento em que a mensagem foi emitida (por exemplo, chuva, jogo de futebol, show de uma banda, etc) e conhecimento sobre o tópico abordado na mensagem (nomes de jogadores, notícias atuais, etc.). A Figura 1 apresenta exemplos de interpretações diferentes que uma mensagem pode produzir durante a segunda etapa da comunicação, dependendo do contexto.

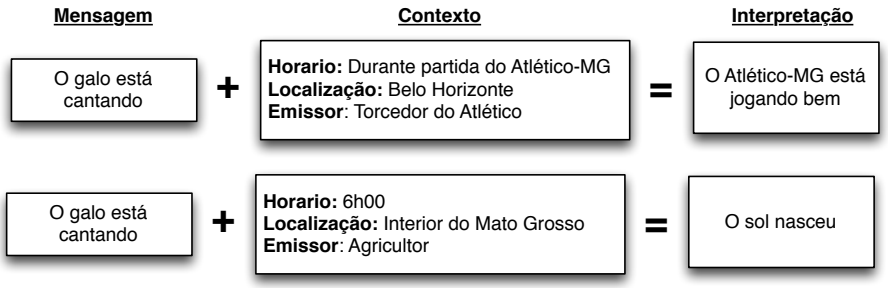

Figure 1. Uma mesma mensagem pode ter duas interpretações diferentes dependendo do contexto

A teoria de relevância de Sperber e Wilson (S&W) [Sperber and Wilson 1986] modela como essas informações contextuais serão utilizadas para a interpretação da mensagem. Ela define um espaço cognitivo pessoal, onde cada indivíduo irá buscar informações contextuais para auxiliar na interpretação de mensagens. O espaço cognitivo de uma pessoa é moldado pelos seus interesses e fontes de informação. Quando duas pessoas se comunicam, segundo S&W, definem um espaço cognitivo mútuo, onde se encontra a interseção das informações contextuais de ambos.

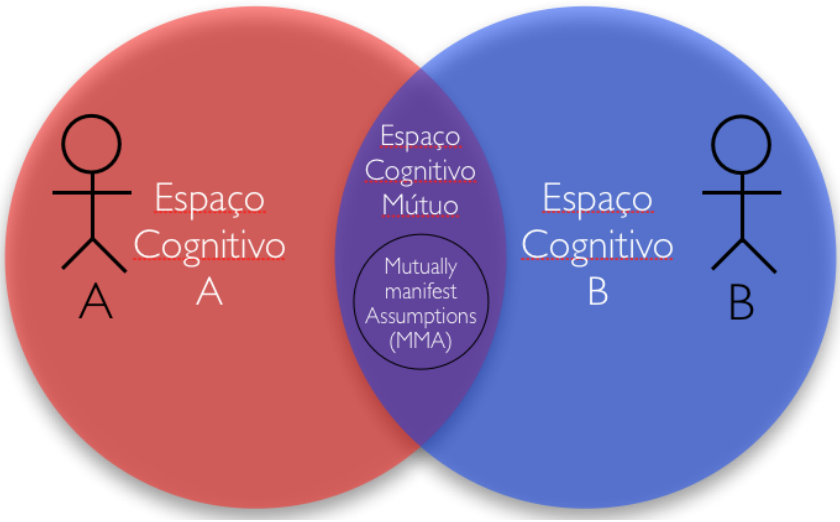

Figure 2. Uma representação do espaço cognitivo entre duas pessoas A e B

Os humanos conseguem aliar sua habilidade interpretativa com seu espaço cognitivo para compreender informações subentendidas e ironia em mensagens. Além disso, inconscientemente, utilizamos essa habilidade para comprimir as mensagens e otimizar a utilização do canal de comunicação. Por exemplo, se uma pessoa deseja perguntar para um são-paulino se ele vai assistir ao jogo no estádio, ela poderia dizer simplesmente: "Você vai ao jogo?" ao invés de: "Você vai ao jogo do São Paulo, hoje, no Morumbi?". Sabendo que o interlocutor torce para o São Paulo, o emissor supõe que ele saiba quando e onde ocorrerá a partida. É natural evitar inserir na mensagem informações redundantes, ou seja, que os dois envolvidos na comunicação já saibam. Esse fenômeno é conhecido, em S&W, como *Mutually Manifest Assumptions* (MMA). Em uma comunicação bem sucedida, o MMA está contido no espaço cognitivo mútuo, ou seja, a suposição que o emissor fez estava correta (Figura 2).

Os MMA são muito comuns em mensagens de microblog (i.e. Twitter) que possuem uma restrição rigorosa de tamanho[1]. Nesses ambientes, os usuários são, muitas vezes, forçados a deixar o máximo de informação implícita para economizar espaço e digitar o texto mais concisamente. Dessa forma, argumentamos que ao utilizar apenas o texto para extrair informações dessas mensagens, grande parte do conteúdo semântico é desconsiderado. São necessárias estratégias que simulem o espaço cognitivo do usuário e suas interações para que possamos entender o real significado que o emissor gostaria de transmitir. No problema da desambiguação, supomos que toda vez que um emissor utiliza uma referência ambígua a uma entidade, ele espera que seus interlocutores sejam capazes de desambiguar utilizando informações contextuais. Nas próximas subseções, são propostas três atributos contextuais que podem ser obtidos utilizando apenas informações do fluxo de mensagens e que são relevantes para a solução do problema: histórico do usuário, localidade temporal e localidade social.

4.1. Histórico de Usuário

No Twitter, quando um usuário segue outro, ele começa a receber todas as mensagens que esse usuário posta na mídia social. Dessa forma, os gostos e interesses (i.e. o espaço cognitivo) de um usuário são assimilados rapidamente por seus seguidores. Da mesma forma, o emissor da mensagem espera que seus seguidores o conheçam suficientemente bem para entender as informações implícitas nesse texto.

Para simular esse comportamento dos usuários, coletamos um fluxo de mensagens S por um longo periodo de tempo (no caso, aproximadamente, 1 ano) utilizando um conjunto de palavras W. Todas as palavras em W estão relacionadas a um conjunto de entidades $E = \{e_1, e_2, ..., e_n\}$ correlacionadas entre si (por exemplo, times em um campeonato de futebol, participantes do Big Brother Brasil, personagens de novela, etc.). Incluímos em W referências ambíguas às entidades E.

Para cada usuário u_i que emitiu pelo menos uma mensagem em S, obtemos a porcentagem H_i das palavras do conjunto W que foram utilizadas por ele durante o período coletado. Consideramos que, quanto maior o valor de H_i, mais amplo será o espaço cognitivo de u_i sobre o conjunto de entidades E. Consequentemente, usuários com H_i alto, ao utilizarem uma palavra w_j ambígua, provavelmente estarão se referindo a uma entidade em E. Por exemplo, se utilizarmos como E todos os times de futebol da primeira divisão do campeonato brasileiro. Pela nossa proposta, um usuário com $H_i = 0,9$ é um especialista em futebol e comenta diversos assuntos relacionados a esse assunto. Logo, quando ele utiliza uma palavra ambígua como "internacional" sabemos com uma alta confiança que ele se refere à entidade "Sport Club Internacional".

4.2. Localidade Temporal da Mensagem

Outra característica importante de redes sociais, em especial o Twitter, que modelamos neste trabalho, é a localidade temporal das mensagens. Em geral, as mensagens são influenciadas por eventos no mundo real (por exemplo, morte de famosos, partidas de futebol, etc.) e por usuários influentes (atores, dirigentes esportivos, políticos, etc.).

Para modelar essa característica, utilizamos o histórico (H_i) para definir um grupo de usuários, R, cujos Tweets se referem a uma das entidades em E com uma probabilidade alta. Para isso, colocamos em R os usuários que possuem os k maiores valores

[1]Mensagens do Twitter só podem ter 140 caracteres.

de H_i, sendo k parâmetro. A métrica proposta para a localidade temporal de uma mensagem L^T é a diferença de tempo da mensagem em relação a última emitida por algum usuário do conjunto R. Dessa forma, quanto menor for o L^T de um tweet, mais próximo temporalmente ele estará da mensagem e, portanto, provavelmente, está se referindo a E (emitida por um usuário em R). Na Figura 3, vemos duas mensagens T_1 e T_2. Pela nossa hipótese, é mais provavel T_1 referenciar a um entidade em E porque $L_1^T < L_2^T$.

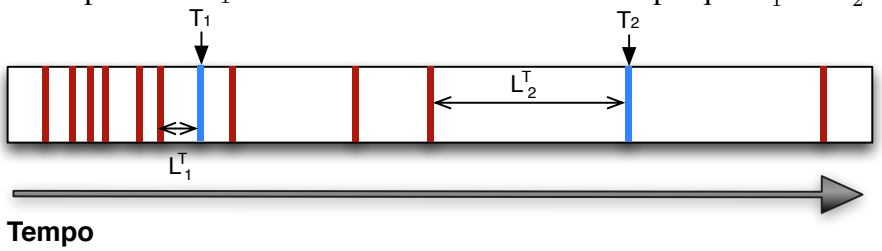

Figure 3. Os traços vermelhos representam mensagens de usuários que pertencem a R. Os traços azuis são mensagens de outros usuários.

A localidade temporal é um atributo importante para modelar, principalmente, tweets de novos usuários, que não possuem grande histórico de mensagens. Além disso, ela complementa a informação do histórico com uma outra fonte de informação, o instante de postagem.

4.3. Localidade Social da Mensagem

O comportamento de usuários em redes sociais é muitas vezes influenciado pelas pessoas com as quais se comunicam. Supomos que toda vez em que há uma comunicação explicita com outro usuário (no caso do Twitter, através de *mentions*, realizados colocando @ antes do nome do usuário), a comunicação foi bem sucedida. Sendo assim, eles possuem espaço cognitivo mútuo não-nulo. Logo, a interpretação desses usuários para um termo ambíguo será, provavelmente, a mesma.

Sendo assim, modelamos essa característica em um grafo não-direcionado cujos vértices representam usuários. As arestas conectam os vértices que mencionam ou são mencionados pelos mesmos usuários. Por exemplo, se o usuário A menciona B e A menciona C, haverá uma aresta entre B e C. O peso das arestas $w(u, v)$ é dado por:

$$w(u, v) = \frac{lift_a(u, v) + lift_p(u, v)}{2} \tag{1}$$

$$lift(a, b) = \frac{P(A \wedge B)}{P(A) * P(B)} \tag{2}$$

O cálculo envolve a métrica de regra de associação $lift$ que mede o grau de "surpresa" provocado pela co-ocorrência de A e B. No caso, o $lift_a$ cobre os casos quando u e v são mencionados pelo mesmo usuário. Já $lift_p$ cobre os casos quando u e v mencionam o mesmo usuário. Esse grafo possui a propriedade de conectar com peso maior usuários semelhantes [Guerra et al. 2011a].

Em seguida, esse grafo é transformado em uma cadeia de Markov, na qual as arestas recebem $w'(u, v) = w(u, v)/\sum w(u, i)$. Nessa cadeia de Markov, calculamos a probabilidade P_u de se chegar a um um vértice qualquer a partir dos usuários no conjunto R (os k maiores H_i) utilizando *random walk* [Spitzer 2001]. Consequentemente, quanto

maior for o P_u, maior será a proximidade desse usuário em relação aos usuários em R. O valor da localidade social de um tweet T_s é igual ao P_u do emissor do tweet.

Essa métrica adiciona informações de rede para complementar o histórico de usuários. Dessa forma, muitos usuários com $0,4 \leq H_i \leq 0,7$ podem ter localidade social alta por estarem próximos e serem semelhantes aos usuários em R.

5. Estudo de Caso e Resultados

Nesta seção, são analisados os resultados dos três atributos propostos em um cenário real de dados coletados do Twitter. Nossa base consiste em cerca de 33,5 milhões de Tweets coletados durante 2012. Nosso conjunto de entidades E é composto por todos os 20 times da série A do Campeonato Brasileiro de Futebol. Para coletar essa base, utilizamos 26 palavras-chave (dentre elas, 10 fortemente ambíguas).

Para validação dos experimentos descritos a seguir, utilizamos três amostras aleatórias dessa base durante os meses de julho a dezembro de 2012. Cada amostra contém 600 tweets com uma referência ambígua a uma das entidades em E. No caso, utilizamos as palavras "galo", "internacional" e "são paulo", que potencialmente se descrevem às respectivas entidades "Clube Atlético Mineiro", "Sport Club Internacional " e "São Paulo Futebol Clube". Anotamos manualmente cada tweet nessa base determinando se a mensagem se refere ao time de futebol ou não. Como *baseline*, utilizamos um classificador associativo sob-demanda (*lazy associative classifier*, LAC) [Veloso et al. 2006]. Utilizamos os 100 primeiros exemplos, em ordem cronológica, do conjunto de tweets rotulados para treinar o classificador e o restante é usado como teste. Como é esperado que o vocabulario mude ao longo do tempo, essa estratégia simula as mudanças de vocabulário comuns em fluxos de dados da mídia social. A seguir, mostramos a eficácia dos modelos contextuais, utilizando curvas ROC (*receiver operating characteristic*), AUC-ROC (*area under ROC curves*) e medidas F1.

5.1. Histórico do Usuário

A Figura 4 apresenta as curvas ROC obtidas para cada um dos conjuntos de teste são mostradas. Percebe-se que várias hipóteses feitas na seção anterior se confirmam nesse experimento. Primeiramente, é fácil notar que os trechos iniciais das curvas são fortemente inclinados (principalmente no dataset "galo"). Isso mostra que, para valores altos de H_i, poucos falsos-positivos são adicionados. Ou seja, usuários com H_i elevado postam pouco conteúdo que não se refere à entidade em questão. Dessa forma, confirmamos a possibilidade de utilizar os H_i mais elevados como "referência" para as outras métricas.

5.2. Localidade Temporal

A Figura 5 traz as curvas ROC para a localidade temporal da mensagem. A eficácia do método foi pior do que no caso anterior, no entanto, essa técnica tem o benefício de não depender diretamente do usuário emissor, apenas do instante em que a mensagem foi emitida. A curva apresenta uma característica diferente: vemos que os últimos pontos da curva apresentam um padrão semelhante ao dos trechos iniciais no histórico do usuário. Isso mostra que um L_T alto é um forte indicador de que a mensagem não se refere à entidade.

5.3. Localidade Social

Para a localidade social, o espaço de amostragem teve que ser reduzido para tweets que ocorreram em novembro, pois gerar o grafo é computacionalmente caro para muitos

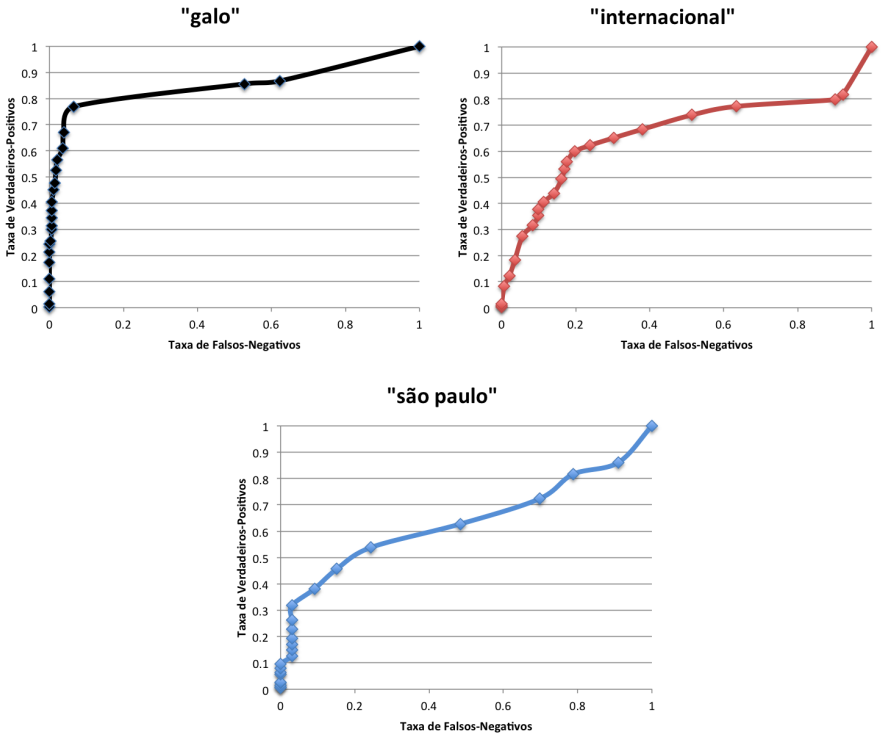

Figure 4. Curvas ROC geradas pelo histórico do usuário

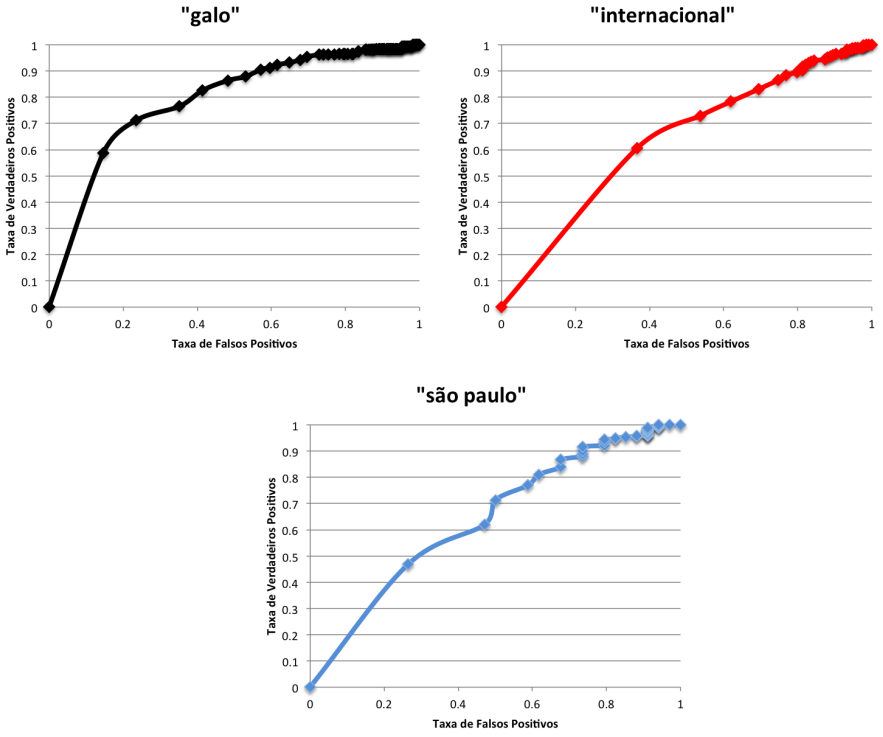

Figure 5. Curvas ROC geradas pela localidade temporal

usuários. No entanto, esse grafo só precisa ser gerado uma vez. Nesse experimento, amostramos mil tweets de um universo de 50 mil (base "galo2"). Com os restantes, geramos o grafo e calculamos o valor de L_S para cada um.

Os resultados se encontram na Tabela 1. Desconsiderando os vértices que não estão no mesmo componente conexo, a localidade social gera um AUC muito alto (0.988). Ou seja, quando o emissor da mensagem está em um mesmo componente conexo que um vértice em R, conseguimos afirmar com alta certeza se o tweet se refere à entidade desejada. No entanto, ao considerar os vértices que não estavam no mesmo componente conexo, o AUC cai bastante (0.826). Uma idéia para tratar esse problema é utilizar a localidade temporal do tweet.

Table 1. Resultados para o modelo de localidade social.

base	Localidade Social (grafo completo)		Localidade Social (desconsiderando vertices não atingíveis pelos usuários em R)		Baseline textual (LAC)
	AUC-ROC	F1	AUC-ROC	F1	F1
galo2	0.826	0.521	0.988	**0.89**	0.891

5.4. Comparação entre os modelos

Os resultados (Tabela 2) mostram que superamos o *baseline* para as referências ambíguas "galo" e "internacional". Na primeira, o histórico obteve resultado superior ao da localidade. Isso mostra que usuários com espaço cognitivo alto utilizam "galo" apenas para se referir ao time. Vemos que no caso de "internacional" o fator temporal é melhor para desambiguar do que as características do usuário. Para a referência "são paulo" obtivemos resultados inferiores ao baseline, principalmente porque a base é fortemente desbalanceada, mostrando que o LAC é mais robusto ao desbalanceamento do que os modelos propostos.

Table 2. Resultados para os modelos de histórico do usuário e localidade temporal.

base	Histórico do Usuário		Localidade Temporal		Baseline textual (LAC)
	AUC-ROC	F1	AUC-ROC	F1	F1
galo	0.904	**0.977**	0.887	0.708	0.893
internacional	0.952	0.86	0.746	**0.875**	0.861
são paulo	0.927	0.9	0.832	**0.96**	0.981

6. Conclusão

Neste trabalho, apresentamos três modelos para extração de informações contextuais de Tweets. Demonstramos que esses modelos estão fortemente correlacionados com o problema da ambiguidade, que esses modelos são invariantes e não requerem atualizações ou intervenções humanas ao longo do tempo para serem mantidos. Finalmente, mostramos que os modelos contextuais relacionados ao usuário (histórico e localidade social) podem ser usados para encontrar pessoas que só comentam sobre o conjunto de entidades E. O modelo temporal tem resultados piores, mas é uma alternativa para mensagens emitidas por usuários com pouco histórico ou que não estão no componente conexo do grafo da localidade social. Além de depender pouco do texto, os modelos são gerais.

Como trabalho futuro, combinaremos os atributos propostos para melhorar o desempenho da desambiguação. Alem disso, pretendemos aplicar essa abordagem para outros problemas de NLP como identificação de entidades e *tagging* de palavras.

References

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). DBpedia - A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3).

Cucerzan, S. (2007). Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 708–716, Prague, Czech Republic. Association for Computational Linguistics.

Davis, A., Veloso, A., da Silva, A. S., Laender, A. H. F., and Meira Jr, W. (2012). Named Entity Disambiguation in Streaming Data. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 815–824.

Guerra, P. H. C., i c Cerf, L., Porto, T. C., Veloso, A., Meira Jr, W., and Almeida, V. i. l. A. F. (2011a). Exploiting Temporal Locality to Determine User Bias in Microblogging Platforms. *JIDM*, 2(3):273–288.

Guerra, P. H. C., Veloso, A., Meira Jr., W., and Almeida, V. (2011b). From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In *KDD '11*, San Diego, CA. ACM Request Permissions.

Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., and Weikum, G. (2011). Robust disambiguation of named entities in text. In *EMNLP '11: Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nguyen, K., Pham, C., Tran, D. A., and Zhang, F. (2011). Preserving Social Locality in Data Replication for Online Social Networks. In *31st IEEE International Conference on Distributed Computing Systems Workshops (ICDCS 2011 Workshops), 20-24 June 2011, Minneapolis, Minnesota, USA*, pages 129–133.

Sperber, D. and Wilson, D. (1986). *Relevance: communication and cognition*. Harvard University Press, Cambridge, MA, USA.

Spitzer, F. (2001). *Principles of random walk*, volume 34. Springer Verlag.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: a core of semantic knowledge. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*. ACM.

Veloso, A., Meira Jr., W., and Zaki, M. J. (2006). Lazy Associative Classification. In *ICDM '06: Proceedings of the Sixth International Conference on Data Mining*. IEEE Computer Society.

Wang, C., Chakrabarti, K., Cheng, T., and Chaudhuri, S. (2012). Targeted disambiguation of ad-hoc, homogeneous sets of named entities. In *WWW '12: Proceedings of the 21st international conference on World Wide Web*. ACM.

Yosef, M. A., Hoffart, J., Bordino, I., Spaniol, M., and Weikum, G. (2011). AIDA: An Online Tool for Accurate Disambiguation of Named Entities in Text and Tables. *PVLDB*, 4(12):1450–1453.

Yus, F. (2011). *Cyberpragmatics Internet-mediated communication in context Pragmatics & Beyond New Series 2011*.