ANÁLISE DE TRABALHOS SIMILARES COM SOLUÇÕES BASEADAS EM PART-OF-SPEECH TAGGING E DESAMBIGUAÇÃO

Pablo Frederico Oliveira Thiele
pablothiele@gmail.com

RESUMO
Este estudo procura mostrar métodos de avaliação, etiquetagem e desambiguação que podem ser utilizados como base para uma proposta de um novo desambiguador. Para tanto, foi realizado um estudo das principais técnicas de etiquetagem, no inglês Part-Of-Speech Tagging as quais foram definidas, culminando com a apresentação de um comparativo tabulado entre estas demonstrando suas características, assim como suas capacidades máximas observadas durante seus trabalhos de pesquisa.

1 - **INTRODUÇÃO**

Atualmente as tecnologias de Processamento de Linguagem Natural (PLN) estão sendo utilizadas em análises de enormes quantidades de dados. Com o advento das novas mídias e a adoção em massa das redes sociais, o fluxo de informações geradas a cada segundo é o maior da história. Segundo o SINTEF, 90% de todas as informações hoje armazenadas no planeta foram geradas nos últimos dois anos [1].

Tendo em mente a necessidade de velocidade de processamento necessário no caso de utilizarmos uma enorme quantidade de informação no procedimento, devemos também, manter a meta de obter a melhor taxa de acerto possível. Sua função, explorada nesse caso é a capacidade de observar e catalogar as palavras de um texto de acordo com suas funções morfossintáticas. O nome comumente dado a este processo é o de POST (Part-Of-Speech Tagging).

Dentro do contexto Part-Of-Speech (POS) encontra-se a função de processar e identificar um grupo de palavras agrupando-as em tipos pré-definidos. Este agrupamento pode ocorrer em razão sintática, morfológica ou morfossintática.

Um dos maiores desafios encontrado em PLN é o problema da ambiguidade. Esta situação que ocorre nas mais diversas etapas do processamento de linguagem natural é complexa, devido à necessidade de que a aplicação processadora tenha conhecimentos abrangentes que possam ser utilizados como ferramentas que colaborem no intuito de realizar as escolhas mais corretas.

De acordo com Silva [2], a ambiguidade é uma das maiores dificuldades a serem devidamente administradas nos processos de linguagem natural. Suas variações podem ser do tipo que representa ambiguidade semântica, sendo esta mitigada com um conhecimento do mundo real do seu redor. Há também as ambiguidades de baixo nível, sendo estas exemplificadas como o reconhecimento correto do ponto final em uma frase, que pode ser tanto para marcar o final desta ou simplesmente indicar a presença de uma palavra que representa uma abreviatura.

Na literatura podem ser encontradas propostas de desambiguação que entremeiam os processos de etiquetação dos textos processados. Citando exemplos, temos os trabalhos Aduriz e Illarraza [3], Brill [4], Giménez e Màrquez [5] e Segond et.al.[7]. Basicamente cada proposta utiliza-se de um horizonte único, alguns trabalham com a noção de utilizar métodos estritamente probabilísticos no momento de definir as desambiguações adequadas. No entanto outros exemplos dão conta de utilizarem uma forma gramatical, criando regras específicas que são utilizadas como guias no momento da catalogação das palavras.

Verifica-se também que nenhuma das propostas encontradas na literatura faz uso de anotações advindas de estruturas *Multi-Terminal Multi-valued Decision Diagrams* (MTMDD) estruturas essas que permitem o uso de grande quantidade de dados de pesquisa, em formato de dicionários, por vezes multilíngue para a classificação de palavras de uma maneira extremamente rápida [6].

O objetivo geral deste estudo é estabelecer uma visão geral do processo de etiquetagem de palavras, compreender seus diversos tipos além de observar como cada trabalho, dentro de seu formato singular trata seus problemas. Consequentemente será possível, utilizando o WAGGER como etiquetador principal, realizar a desambiguação de classes gramaticais de forma automática, em textos de língua portuguesa. A seguir um breve resumo de como isso será produzido.

A primeira etapa consistirá em criar ferramentas próprias para ajustar um conjunto de *corpus* que será utilizado para a avaliação. Dentre diversas opções existentes, este trabalho irá focar exclusivamente no Mac-Morpho[1]. Aproveitando-se a característica do Mac-Morpho de ser um corpus revisado com etiquetagem correta, a avaliação da qualidade do *parser* será medida através

[1] Mac-Morpho é um corpus formado a partir de nóticias em português brasileiro anotado com POS tags [8].

do seu comparativo com o texto previamente anotado. O intuito do processo de melhoria contínua é adicionar novas funcionalidades e escopo aos poucos, assim para cada novo processo criado um melhor desempenho do desambiguador é esperado.

Como soluções funcionais, os desambiguadores serão testados com os *corpora* previamente definidos. Passado o período de testes, obtendo valores adequados para cada tipo de abordagem, é o momento para novos ajustes e melhorias eventuais a fim de gerar uma proposta que consiga trazer uma desambiguação adequada para textos que foram previamente anotados através de dicionários MTMDD existentes.

2 – TRABALHOS SIMILARES

Os trabalhos similares buscados foram aqueles que conseguiram em seus diversos métodos bons resultados em relação às técnicas mais conhecidas. Ainda, para ter uma boa visão das possibilidades que essa área de estudo apresenta, foram escolhidos trabalhos que usem maneiras de POS *tagging* baseadas em probabilidades ou que utilizem regras de desambiguação.

Na área de desambiguação através de regras, o trabalho de Brill [4] pode ser considerado um artigo base. Essa proposta descreve um algoritmo que é capaz de realizar aprendizagem sem supervisão, aprendendo assim a partir de corpus não anotados manualmente. Neste trabalho ele também apresenta uma integração das possibilidades de treinamento para a obtenção de regras, combinando algoritmos de treinamento supervisionado e não supervisionado criando assim um *tagger* de alto desempenho com uma necessidade pequena de textos pré-anotados.

Seguindo na linha de desambiguação baseada em regras, o trabalho de Aduriz e Illarraza [3], aparece como um trabalho onde uma análise de ambiguidades morfossintáticas é realizada, sendo focado especialmente no idioma basco. Essa proposta utiliza como *parser* o já comentado *Constraint Grammar* (CG) [25] que já foi diversas vezes utilizado na criação de gramáticas para vários idiomas diferentes. Sua colaboração foi apresentar um resultado de melhoria nas tarefas de identificação e ambiguidades em funções sintáticas e morfossintáticas utilizando uma análise superficial. Para em seguida realizar as desambiguações através das mil regras geradas para este fim.

Exemplo de um representante da linha de pesquisa de *taggers* estatísticos, o trabalho de Giménez e Màrquez [5] apresenta uma proposta de POS tagger baseado em *Support Vector Machines* chamado de *SVMTool*. Essa ferramenta se apresenta como uma opção simples, flexível, e eficiente para as necessidades atuais do processamento de linguagem natural. De acordo com os autores a *SVMTool* simples de utilizar necessitando de poucos parâmetros para funcionar em seu formato na linguagem *Pearl*. O contexto a ser adotado na ferramenta também é passível de ajuste, definindo tamanhos dos N-gramas (bigramas, trigramas, etc.), podendo ajustar também o tempo de *tagging* a ser executado. Outra vantagem apresentada no trabalho é que a ferramenta tende a se comportar bem não importando o idioma utilizado. Nos testes, utilizando tanto o inglês quanto o espanhol foi possível observar marcas consideráveis nas taxas globais, (96,16% e 96,89% respectivamente). Para tanto além de uma etapa de aprendizagem não supervisionada de textos no idioma pretendido, é necessário adicionar às suas configurações dicionários morfossintáticos compatíveis com estas linguagens.

Mantendo as propostas baseadas em probabilidades, o trabalho de Segond et.al.[7] apresenta um experimento que através de HMM consiste em uma formatação clássica de um *tagger* deste estilo. A preparação de dados se dá com a obtenção de todas as *tags* semânticas possíveis obtidas na *WordNet*[2]. Em seguida, tendo como origem o *Brown Corpus*, foi gerado um *corpus* de treinamento e um corpus para os testes propriamente. Nos testes o modelo HMM utilizado apenas era capaz de interpretar bigramas, isto é cada palavra apenas leva em consideração a palavra imediatamente anterior. Por fim realizado o processamento do *corpus* de treino a fim de realizar os ajustes necessários no algoritmo de *tagging*, foi processado o *corpus* de teste. Na sequência foi realizada uma comparação das *tags* encontradas nos no processamento, com as já existentes no conjunto de palavras do treinamento inicial que já haviam sido etiquetadas manualmente. Três testes distintos foram realizados, conseguindo uma taxa máxima de acerto global de 89%. Como possibilidade de melhoria o trabalho conclui que adicionar a capacidade de consulta a dicionários sintáticos poderia incrementar o desempenho obtido através da formatação clássica executada.

Como exemplo a ser destacado, com um formato já ajustado e adequado à utilização do idioma português encontra-se a tese de Domingues [10] que propõe uma abordagem completa para o desenvolvimento de um etiquetador de alta acurácia para o português do Brasil. Esse estudo

[2] WordNet é um grande banco de dados com dados léxicos no idioma inglês. Cada um expressando um conceito, substantivos, verbos, adjetivos e advérbios são agrupados em conjuntos de sinônimos cognitivos, os *synsets* [9].

exploratório apresenta solução que foi idealizada como uma visão híbrida combinando etiquetagem probabilística e etiquetagem baseada em regras. Foram utilizados quatro versões dos seguintes corpora CETENFolha, Bosque CF 7.4, Mac-Morpho e Selva Científica. A solução de *tagging* foi utilizar ferramentas *open source* já existentes, entre elas o gerador de regras automatizado μ-TBL (*Micro transformation-based learning*) e o TreeTagger como solução de tagging. Sua solução final apresenta diversas fases de processamentos, que começa com a tokenização do texto, passando pelo *parser* estatístico. Após essa etapa ocorre a consulta por nomes próprios no texto e por fim a etiquetagem baseada em regras. Sendo esta última etapa diferenciada, pois se utiliza de três grupos de regras. Todo esse processo permitiu que os experimentos com melhor desempenho atingissem uma taxa de acerto global superior a 98%.

3 – CONCLUSÃO

Para sintetizar as informações coletadas nos trabalhos similares apresentados, as principais características relevantes para este trabalho foram organizadas e são apresentadas na Tabela 1. Esta análise, todavia não possui um caráter de escolha entre as propostas a seguir apresentadas. Servem, no entanto para obter um entendimento sistemático das opções existentes para o problema. Desta forma a contínua observação dos trabalhos já utilizando as ferramentas adequadas para este problema, colabora na compreensão ao produzir subsídios necessários para a escolha de um método de desambiguação próprio. Esse método busca trazer uma nova contribuição ao, aplicar soluções realizadas em outros idiomas tornando-o útil para textos em português, ou ainda conseguir uma mescla destas possibilidades.

Tabela 1 - Comparação entre trabalhos similares encontrados na bibliografia

Identificação da Proposta	**Tipo de *tagging***	**Máxima acurácia obtida em testes**	**Em qual idioma é aplicada**	**Tipo de desambiguação utilizado**
Aduriz e Illarraza [3]	Baseado em regras, usando uma gramática própria.	97.51%	Basco	Regras de desambiguação
Brill [4]	Baseado em regras, com aprendizagem sem supervisão. (Transformation-Based Learning)	96%	Inglês	Regras de desambiguação
Giménez e Màrquez [5]	Utilizando *Support Vector Machines* (SVM)	97.16%	Inglês, Espanhol	Estatístico
Segond et.al.[7]	Baseado em HMM	89%	Inglês	Estatístico
Domingues [10]	Processo híbrido baseado em probabilidade e regras	98,30%	Português do Brasil	Estatístico/Regras

5 – REFERÊNCIAS BIBLIOGRÁFICAS

[1] SINTEF, BIG *Data, for better or worse: 90% of world's data generated over last two years, In ScienceDaily. Disponível em: <http://www.sciencedaily.com/releases/ 2013/05/1305220 85217.htm>. Acesso em: 12 de dezembro de 2013.*

[2] SILVA, JOÃO RICARDO, *SHALLOW Processing of Portuguese: From Sentence Chunking to Nominal Lemmatization*, In Master's thesis, Lisboa University, Portugal. 2007

[3] ADURIZ ITZIAR, ILLARRAZA ARANTA DÍAS DE, *Morphosyntactic disambiguation and shallow parsing in computational processing of Basque*. In Inquiries into the lexicon-syntax relations in Basque / Bernard Oyharçabal (aut.), 2004, ISBN 84-8373-580-6, pp. 1-23.

[4] BRILL ERIC, Unsupervised *Learning of Disambiguation Rules for Part of Speech Tagging*. In Natural Language Processing Using Very Large Corpora. Kluwer Academic Press, 1995, pp. 1-13.

[5] GIMÉNEZ JESÚS, MÀRQUEZ LLUÍS, *Svmtool: A general pos tagger generator based on support vector machines*. In Proceedings of the 4th International Conference on Language Resources and Evaluation, 2004, pp. 43-46.

[6] FERNANDES PAULO, LOPES LUCELENE, PROLO CARLOS AUGUSTO, SALES AFONSO, VIEIRA RENATA, *A Fast, Memory Efficient, Scalable and Multilingual Dictionary Retriever*. In LREC, 2012, pp. 2520-2524.

[7] SEGOND FREDERIQUE, SCHILLER ANNE, GREFENSTETTE GREGORY, CHANOD JEAN-PIERRE, *An Experiment In Semantic Tagging Using Hidden Markov Model Tagging*. In ACL/EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, 1997, pp. 78-81.

[8] NILC - USP, MAC-MORPHO. Disponível em < http://www.nilc.icmc.usp.br/macmorpho/> Acesso em 19 de agosto de 2014.

[9] PRINCETON UNIVERSITY, "About WordNet." Wordnet. Disponível em <http://wordnet. princeton.edu> Acesso em 12 de agosto de 2014.

[10] DOMINGUES, MIRIAM LÚCIA CAMPOS SERRA. Abordagem para o desenvolvimento de um etiquetador de alta acurácia para o Português do Brasil. 2011. 140 f. Tese (Doutorado) – Universidade Federal do Pará, Instituto de Tecnologia, Belém, 2011. Programa de Pós-Graduação em Engenharia Elétrica.