Rotulação Semântica Automática de Sentenças para a FrameNet

William Paulo Ducca Fernandes[1]

[1]Faculdade de Letras
Universidade Federal de Juiz de Fora (UFJF) – Juiz de Fora – MG – Brazil

william.ducca.fernandes@ice.ufjf.br

Abstract. *This work aims to implement an automatic semantic labeler of sentences for FrameNet – semantic network which has been, for more than 10 years, developed for English. The problem of semantic labeling of sentences is expressed as a classification problem and the Support Vector Machines model is used as classifier. Experimental results show the labeling task presents good results.*

Resumo. *Este trabalho objetiva implementar um rotulador semântico automático de sentenças para a FrameNet – rede semântica que vem sendo, há mais de dez anos, desenvolvida para o Inglês. O problema de rotulação semântica de sentenças é expresso como um problema de classificação e como classificador é utilizado o modelo de Máquinas de Vetores Suporte. Os resultados experimentais mostram que a tarefa de rotulação apresenta bons resultados.*

1. Introdução

A FrameNet é um projeto sediado no Instituto Internacional de Ciência da Computação (ICSI) na Universidade da Califórnia em Berkeley. Seu objetivo é produzir descrições de unidades lexicais da língua inglesa, através de procedimentos automáticos e manuais de anotação.

Uma característica importante da FrameNet é que ela pode ser acoplada a múltiplas realidades lingüísticas, visto que ela é uma rede de estruturas conceituais e, desta forma, pode representar conceitos de qualquer língua.

Uma rede semântica deste tipo é de grande interesse para diversas aplicações de Processamento de Linguagem Natural (PLN), tais como tradução automática, extração de informação, paráfrases e sumarização de textos.

Neste contexto, o trabalho proposto implementa um rotulador semântico automático de sentenças de forma a mostrar que esta rede pode ser usada como base para outras tarefas de PLN.

O problema de rotulação semântica de sentenças é expresso como um problema de classificação. Como classificador, é utilizado o modelo de Máquinas de Vetores Suporte, pois, em geral, ele apresenta alta precisão sem a necessidade de alta qualidade nos dados de treinamento.

2. FrameNet

A FrameNet é um projeto de lexicografia computacional [Baker, Fillmore e Lowe, 1998] que extrai informação das propriedades semântico-sintáticas de unidades lexicais[1] (ULs) de extensos corpora eletrônicos. Isto é feito através de processos de anotação automática e manual de sentenças.

O nome do projeto teve inspiração num projeto similar, WordNet, entretanto é baseado na teoria da Semântica de Frames e, desta forma, está interessado nas redes semânticas de que as ULs participam.

Frames são estruturas conceituais internamente complexas que são definidas em termos dos participantes que as integram (elementos do frame). Por exemplo, o frame **Chegada** é composto por quatro elementos[2]: um TEMA, uma ORIGEM, um CAMINHO e um DESTINO, de forma que o TEMA se move de uma ORIGEM ao longo de um CAMINHO até um DESTINO. A FrameNet então busca pelos usos lingüísticos de ULs como *chegar*, *vir*, *retornar*, *entrar*, *visitar* que realizam este frame e descreve as dependências sintáticas e semânticas destes itens. A figura 1 mostra três exemplos de anotação dos elementos do frame e das ULs.

[$_{\text{TEMA}}$Ricardo] **retornou** [$_{\text{ORIGEM}}$do sítio] na madrugada de segunda.
Ontem [$_{\text{TEMA}}$Delza] **chegou** [$_{\text{DESTINO}}$em casa] cedo [$_{\text{ORIGEM}}$da escola].
[$_{\text{TEMA}}$Uns meninos] **entraram** [$_{\text{DESTINO}}$no parque] [$_{\text{CAMINHO}}$através de um buraco na cerca], depois que ele estava fechado.

Figura 1. Exemplos de anotação dos elementos do frame e das ULs

A FrameNet também descreve as relações semânticas estabelecidas pela própria rede de frames, pois eles se vinculam principalmente através de relações de Herança, Uso, Subframe e Perspectiva. Por exemplo, o frame **Chegada** herda elementos do frame **Transporte** (elementos ORIGEM, DESTINO etc). O frame **Velocidade** usa (pressupõe) o frame **Movimento**. O frame **Julgamento** é um subframe do frame **Processo_criminal** e o frame **Emprego** envolve a perspectiva do EMPREGADOR (recrutamento) e a do EMPREGADO (admissão).

3. Máquinas de Vetores Suporte

Máquinas de Vetores Suporte (SVMs) são uma técnica de aprendizado supervisionado, da área de aprendizado de máquinas, que objetiva realizar tarefas de classificação[3].

O objetivo fundamental desta técnica é a obtenção de um classificador com alto poder de generalização. É necessário que ele tenha capacidade de separar de forma

1 Unidade lexical é uma palavra com determinado significado.

2 O frame **Chegada** é composto por mais elementos. Foram colocados apenas quatro a caráter explicativo.

3 As SVMs também são usadas em tarefas de regressão, mas esse tema não será abordado, visto que não é o objetivo deste trabalho.

correta o conjunto de dados de treinamento e também que ele seja capaz de classificar corretamente os dados na etapa de teste [Cristianini e Shawe-Taylor, 2000].

O conceito chave por trás das SVMs pode ser descrito em poucas palavras. Dado um conjunto de treinamento que contém pontos pertencentes a duas classes distintas, a SVM traça um hiperplano à partir de certos pontos do treinamento (vetores suporte). Seu objetivo é a separação dos pontos das duas classes e a maximização da margem de separação (distância que separa o hiperplano dos vetores suporte de cada classe).

3.1. Técnica de solução

Embora seja construída uma solução de máxima margem de separação com as SVMs, elas são, em sua formulação, um problema quadrático. Em busca de técnicas que reduzam o tempo de execução quadrático, novos algoritmos têm sido desenvolvidos, motivados no fato de que geralmente só é necessária uma aproximação da margem ótima para se obter boa capacidade de generalização [Leite e Neto, 2007].

Neste contexto, é apresentado o algoritmo de margem incremental (IMA), um algoritmo linear que constrói uma aproximação da solução de máxima margem.

Os resultados produzidos pelo IMA são comparáveis aos das SVMs, o que pode ser visto no trabalho de Leite e Neto (2007).

4. Experimento Computacional

Foi realizado um experimento computacional com o objetivo de verificar a viabilidade da rotulação semântica automática de sentenças.

Esta tarefa de rotulação é expressa como um problema de classificação e o algoritmo escolhido foi o IMA pelo fato de ser linear.

Para cada frame foi treinado um classificador diferente. Esta decisão foi tomada com base na análise do corpus de teste. Na grande maioria das vezes, as sentenças analisadas pertenciam a pelo menos dois frames distintos.

Os dados usados no treinamento e no teste foram obtidos das sentenças geradas pelo processo de anotação da FrameNet, disponíveis no site oficial da FrameNet.

O procedimento de conversão de sentenças para vetores numéricos baseia-se no trabalho de Moschitti, Morarescu e Harabagiu (2003) com duas particularidades: as palavras são lematizadas[4] antes do processo de conversão; e ocorre apenas a extração de informação lexical das sentenças.

Assim cada palavra distinta do conjunto de treinamento representa uma característica distinta. Logo cada sentença *s* tem um vetor associado na forma:

$$\vec{f}_s = \langle f_1^s, f_2^s, ..., f_N^s \rangle \quad (4.1)$$

onde $f_i^s = 1$ apenas se a i-ésima característica aparece na sentença *s*. Caso contrário, $f_i^s = 0$.

O peso de cada característica f_i^s também é computado em cada sentença. Para isso, primeiramente é computada a freqüência de aparecimento da característica f_i na sentença *s*. Depois é medido o número de sentenças N_{f_i} do corpus em que f_i está

4 No processo de lematização, as palavras são convertidas para a sua forma neutra. Por exemplo, a palavra *retornou* é convertida para *retornar*.

presente. Com esses valores em mãos, é obtido o *Inverso da Freqüência da Sentença* (ISF) para a característica f_i da seguinte forma:

$$ISF(f_i) = \log\left(\frac{n_c}{N_{f_i}}\right) \tag{4.2}$$

onde n_c é o número total de sentenças do corpus.

O peso para cada característica é dado por:

$$\omega^s_{f_i} = \frac{r^s_{f_i} \cdot ISF(f_i)}{\left(\sum_{j=1}^{N} [r^s_{f_j} \cdot ISF(f_j)]^2\right)^{1/2}} \tag{4.3}$$

Assim o vetor de pesos associado a cada sentença *s* é:

$$\vec{\omega}_s = <\omega_{f1}{}^s, \omega_{f2}{}^s, ..., \omega_{f3}{}^s> \tag{4.4}$$

4.1. Resultados

No experimento, mediu-se o desempenho do classificador através da precisão da rotulação, do recall e do desempenho combinado *f*. Também foi medido o desempenho médio dos classificadores para cada um dos três tipos de medições.

O treinamento foi realizado com 5.308 sentenças pertencentes a 10 frames das anotações lexicográficas da FrameNet.

Os frames foram escolhidos com base em dois critérios. Primeiramente os que retornaram mais ocorrências no corpus de teste foram priorizados. Depois foram selecionados os que tinham mais ULs anotadas.

O teste foi realizado com 1.257 sentenças das anotações de texto completo da FrameNet que constituem o corpus NTI.

A tabela 1 mostra o desempenho dos classificadores para os 10 frames escolhidos. A precisão varia de 84,25% a 99,68%, enquanto o recall varia de 25,97% a 80,65%. Os resultados também mostram o desempenho médio combinado *f* de 64,90%.

Tabela 1. Resultados do experimento

Nome	Precisão	Recall	*f*
Obtenção	97,06	48,53	64,71
Possessão	95,70	36,84	53,20
Afirmação	84,25	79,80	81,96
Sinal	99,68	66,67	79,90
Revelação	99,05	38,89	55,85
Primeiro_ranque	99,12	35,29	52,05
Pesquisa	95,07	25,97	40,80
Tentativa	97,85	41,30	58,09
Especialização	98,33	80,65	88,61
Invenção	99,12	58,82	73,83
Média	**96,52**	**51,28**	**64,90**

5. Conclusão

Neste trabalho foi implementado um rotulador semântico automático de sentenças para a FrameNet através da utilização do modelo de Máquinas de Vetores Suporte.

Os resultados experimentais mostraram que a tarefa de rotulação apresenta bons resultados – mas que ainda podem ser melhorados. Isto possibilita que esta rede seja usada como base para aplicações de PLN.

5.1. Trabalhos futuros

Com base neste trabalho, três outros podem ser desenvolvidos posteriormente. O primeiro deles envolve a melhoria da tarefa de rotulação, o que possivelmente se dará através da extração de informação sintática das sentenças analisadas.

O segundo envolve a adaptação da aplicação implementada para a FrameNet Brasil, versão da FrameNet para o Português em desenvolvimento na Universidade Federal de Juiz de Fora, sob a liderança da lingüista Margarida Salomão.

O terceiro trabalho envolve a identificação de papéis semânticos, onde os elementos do frame e a UL de cada sentença rotulada deverão ser identificados.

Referências

Baker, C. F.; Fillmore, C. J.; Lowe, J. B. (1998) The Berkeley FrameNet Project. In: Proceedings of COLING-ACL, Montreal, p. 86-90.

Cristianini, N.; Shawe-Taylor, J. (2000) **An Introduction to Support Vector Machines and other kernel-based learning methods**. [S.I.]: Cambridge University Press.

Leite, S. C.; Neto, R. F. (2007) Incremental margin algorithm for large margin classifiers. Neurocomputing, v. 71, p. 1550-1560.

Moschitti, A.; Morarescu, P; Harabagiu, S. M. (2003) Open-domain information extraction via automatic semantic labeling. In: Proceedings of FLAIRS, St. Augustine, p. 397-401.