POS Tagger basado en HMM y SVM

Javier Redolfi

CIII - Centro de Investigación en Informática para la Ingeniería
Universidad Tecnológica Nacional - Facultad Regional Córdoba
Córdoba, Argentina

Curso de Procesamiento del Lenguaje Natural, 2013

Objetivos

- Construcción de un POS Tagger
- Especializado en preguntas
- Con salida probabilística
- Que permita trabajar al revés
  - Que la entrada sean los POS Tags
- Que permita el faltante de features (missing features)

Asunciones

- Alto grado de correlación entre un POS y sus vecinos anteriores y posteriores
- Fuerte estructura de las oraciones a procesar (preguntas)
  - Empiezan y terminan con signos de interrogación
  - En las primeras palabras hay pronombres interrogativos

Solución Propuesta

Modelos Ocultos de Markov

- Usados para reconocimiento temporal de patrones
  - habla
  - dígitos escritos a mano
  - gestos
  - POS Tagging
  - partituras musicales

Solución Propuesta

Modelos Ocultos de Markov para POS Tagging

- Permiten modelar el estado inicial más probable
- Permiten modelar la transición de un POS con sus vecinos
- Permiten modelar la generación de las features dado el POS
- Al necesitar entrenamiento se puede especializar para preguntas
- Tiene salida probabilística
- Permite que las entradas sean los POS Tags (salidas)

HMM

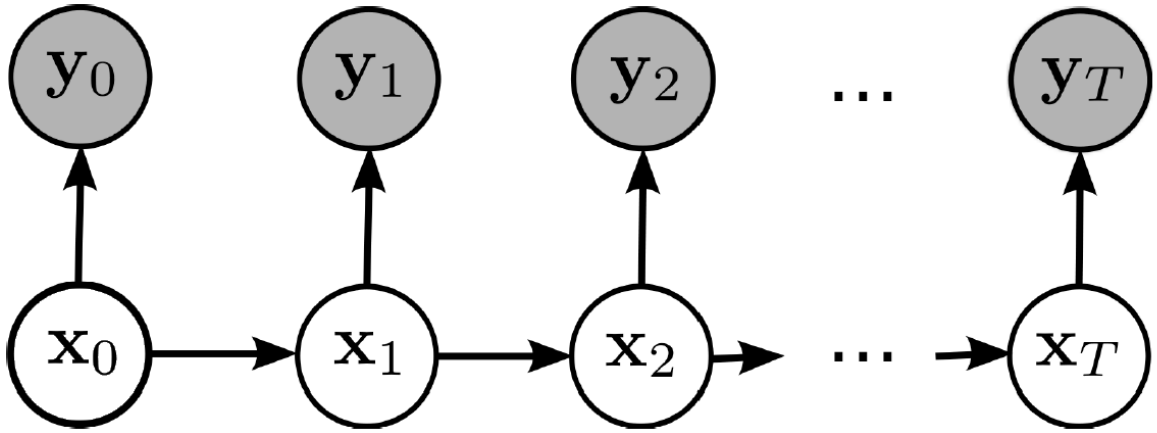

Figura : Modelo Oculto de Markov

- **x** $\rightarrow$ estados ocultos, POS tags
- **y** $\rightarrow$ observaciones, features

HMM

Probabilidades de Transición

probabilidad de pasar de un estado a otro

- cuentas en un set de entrenamiento etiquetado con sus POS Tags

Probabilidades de Emisión

probabilidad de una determinada feature dado el estado

- se necesita un modelo para cada POS Tag
- se entreno un SVM lineal con salida probabilística

Features

- POS Tag
- número, signo de puntuación, mayúsculas
- sufijos de tamaño 1
- sufijos de tamaño 2, los más comunes
- sufijos de tamaño 3, los más comunes
- prefijos de tamaño 2, los más comunes
- prefijos de tamaño 3, los más comunes

Configuración Experimental

Corpus

- 4000 preguntas anotadas
- 90 % para entrenamiento, 10 % para test

Baseline

- Senna POS Tagger
- NLTK HMM POS Tagger

Herramientas

- NLTK, scikit-learn, libsvm, senna

Análisis de Características

Agregado de sufijos como características

pregunta	What	is	the	name	of	the	managing	director	of	...
ground truth	WP	VBZ	DT	NN	IN	DT	JJ	NN	IN	...
POS+REG+suf1	WP	VBZ	DT	NN	IN	DT	NN	NN	IN	...
POS+REG+suf2	WP	VBZ	DT	NN	IN	DT	JJ	NN	IN	...

Agregado de verbos auxiliares como características

pregunta	What	is	the	name	of	a	hotel	in	Indianapolis	...
ground truth	WP	VBZ	DT	NN	IN	DT	NN	IN	NNP	...
POS	WP	VBD	DT	NN	IN	DT	NN	IN	NNP	...
POS+REG+suf2	WP	VBZ	DT	NN	IN	DT	NN	IN	NNP	...

Preguntas terminadas en Punto

pregunta	Name	a	stimulant	.
ground truth	VB	DT	NN	.
POS	VB	DT	NN	?

Resultados

Senna POS Tagger	0.8385
NLTK HMM POS Tagger	0.8195
POSTag	0.8610
POSTag + REG	0.8779
POSTag + REG + suf1	0.8811
POSTag + REG + suf12	0.8798
POSTag + REG + suf123	0.8798
POSTag + REG + suf123 + pre2	0.8808
POSTag + REG + suf123 + pre23	0.8813

Figura : Comparación de la precisión usando diferentes features

Conclusiones

Conclusiones

- Se diseñó y construyó un POS Tagger
- Se cumplieron con todos los objetivos, excepto
- el de manejar características faltantes
- Para superar esto se podría usar ensemble classifiers

Selección de Características

- elegir las más discriminativas
- permite simplificar el problema

Problemas de la Selección de Características

- no interpretable
  - por ejemplo PCA
  - se adapta mejor a muchos datos y features
- interpretable
  - por ejemplo correlación
  - difícil de analizar cuando tenemos muchas clases y features

Uso de otras Características

- palabras funcionales
  - preposiciones
  - verbos auxiliares
- clases de palabras
  - días de la semana
  - meses
  - clustering