SVMTool: A general POS tagger generator based on Support Vector Machines

4th INTERNATIONAL CONFERENCE

ON LANGUAGE RESOURCES AND EVALUATION

Jesús Giménez and Lluís Màrquez

May 26, 2004

TALP Research Center, Universitat Politècnica de Catalunya

Outline

- **Introduction**
  - **Part-of-speech Tagging**
  - **Idea and Motivation**
  - **Learning Framework**
- SVMT tool
- Evaluation
- Conclusions

Introduction

- Part-of-Speech Tagging

> The_DT SVMTool_NNP is_VBZ now_RB being_VBG presented_VBN to_TO NLP_NNP researchers_NNS in_IN Lisbon_NNP ._.

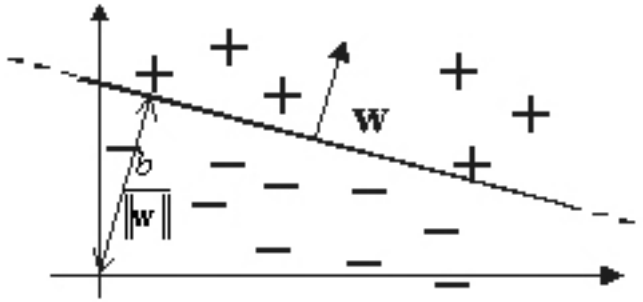
- Brill [Brill, 1995]
- TnT [Brants, 2000]

Idea and Motivation

- Accuracy [in the state-of-the-art]
- Efficiency [both learning and tagging]
- Flexibility [highly customizable]
- Portability [language independent]
- Robustness [against overfitting and on-line mistakes]
- Simplicity [easy to use]

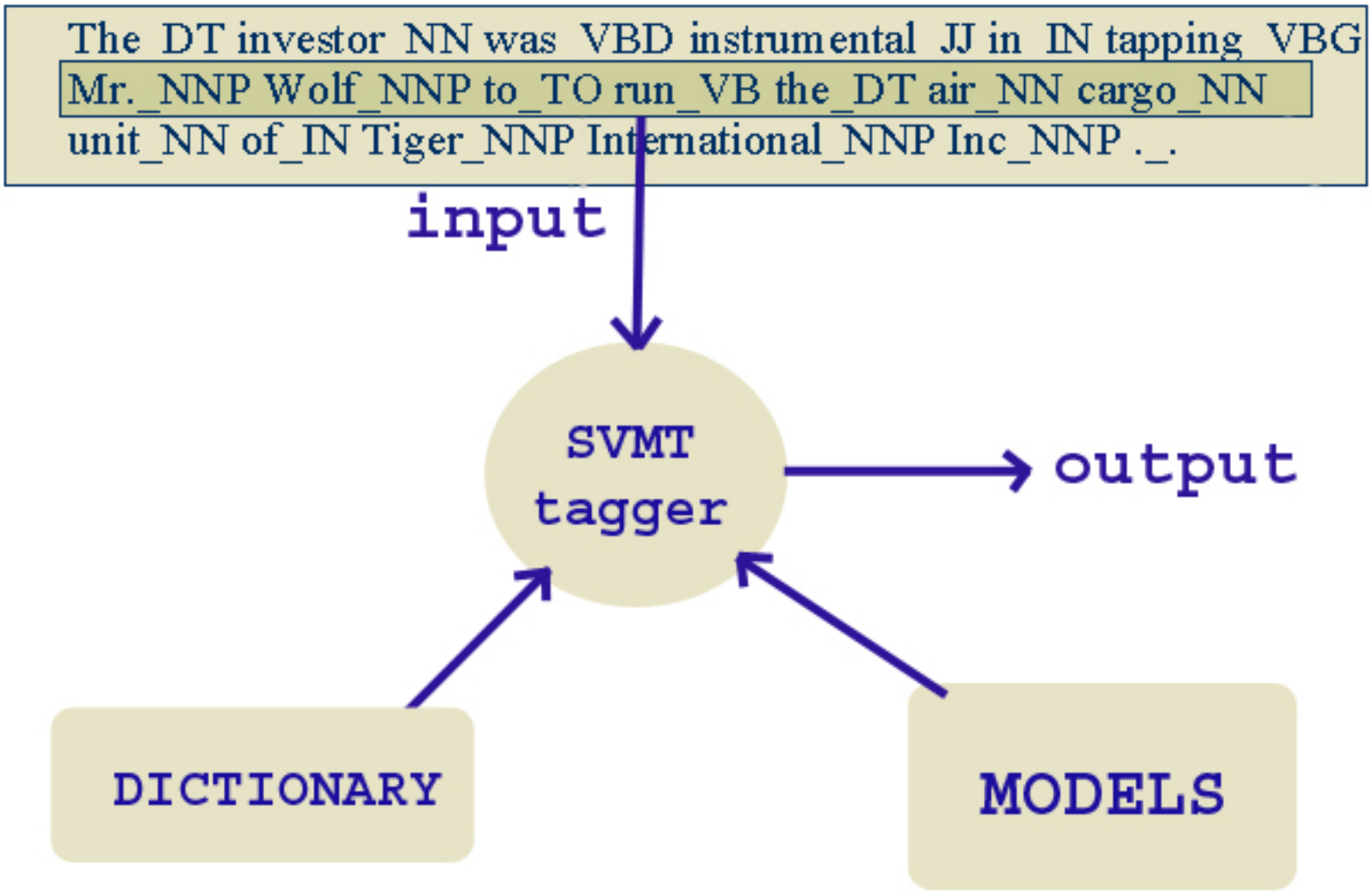
Learning FrameWork

- Support Vector Machines



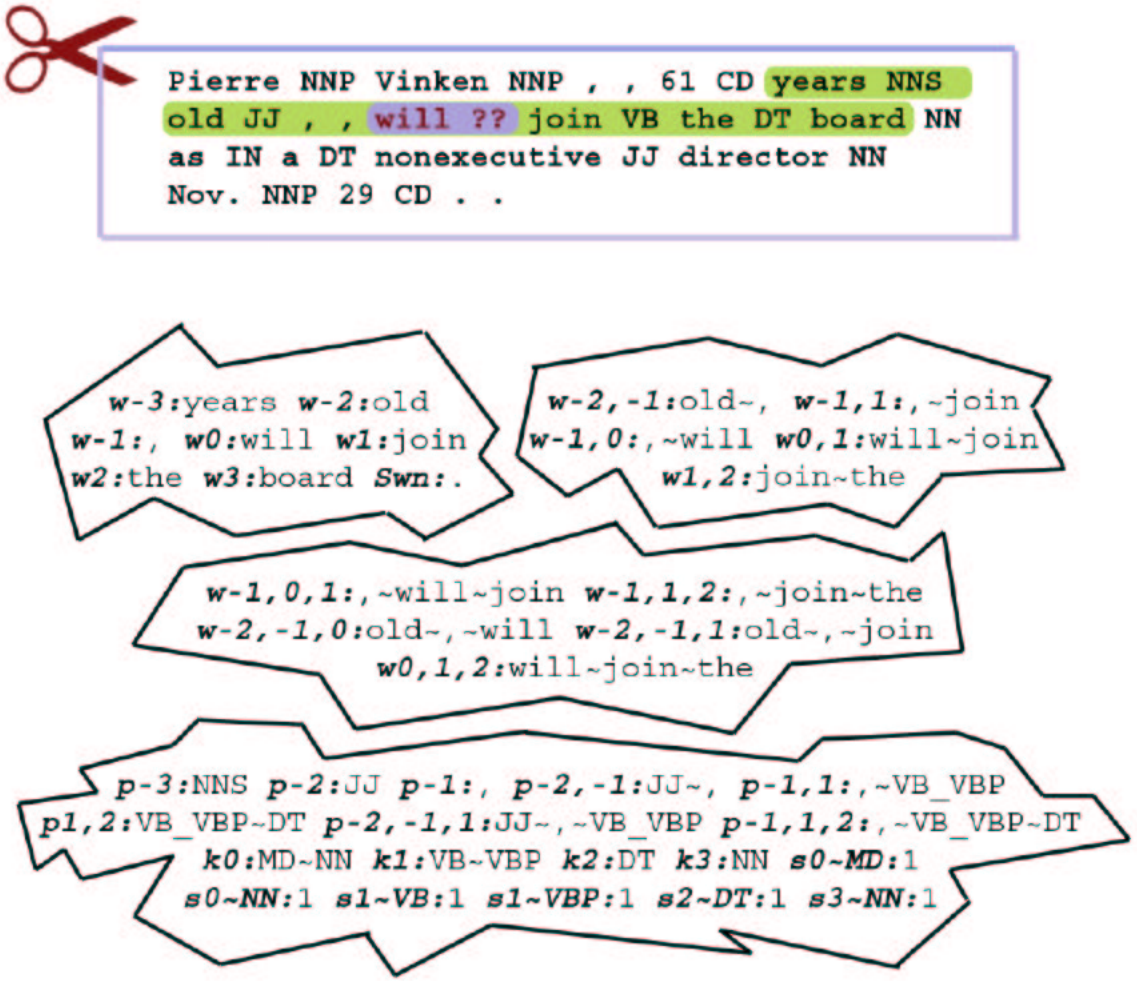
$$h(\mathbf{x}) = \text{sign}\left(\langle \mathbf{w} \cdot \mathbf{x} \rangle + b\right) = \begin{cases} +1 & \text{if } \langle \mathbf{w} \cdot \mathbf{x} \rangle + b > 0 \\ -1 & \text{otherwise} \end{cases}$$

SVMT tool


The_DT investor_NN was_VBD instrumental_JJ in_IN tapping_VBG
Mr._NNP Wolf_NNP to_TO run_VB the_DT air_NN cargo_NN
unit_NN of_IN Tiger_NNP International_NNP Inc_NNP ._.
input
SVMT
tagger
output
DICTIONARY
MODELS

Feature Patterns

word unigrams **word bigrams** **word trigrams**	$w_{-3}, w_{-2}, w_{-1}, w_0, w_{+1}, w_{+2}, w_{+3}$ $(w_{-2}, w_{-1}), (w_{-1}, w_{+1}), (w_{-1}, w_0), (w_0, w_{+1}), (w_{+1}, w_{+2})$ $(w_{-3}, w_{-2}, w_{-1}), (w_{-2}, w_{-1}, w_0), (w_{-2}, w_{-1}, w_{+1}),$ $(w_{-1}, w_0, w_{+1}), (w_{-1}, w_{+1}, w_{+2}), (w_0, w_{+1}, w_{+2})$
POS unigrams **POS bigrams** **POS trigrams**	p_{-3}, p_{-2}, p_{-1} $(p_{-2}, p_{-1}), (p_{-1}, a_{+1}), (a_{+1}, a_{+2})$ $(p_{-3}, p_{-2}, p_{-1}), (p_{-2}, p_{-1}, a_{+1}), (p_{-1}, a_{+1}, a_{+2})$
ambiguity classes	a_0, a_1, a_2, a_3
maybe's	m_0, m_1, m_2, m_3
prefixes	s_1, $s_1 s_2$, $s_1 s_2 s_3$, $s_1 s_2 s_3 s_4$
suffixes	s_n, $s_{n-1} s_n$, $s_{n-2} s_{n-1} s_n$ $s_{n-3} s_{n-2} s_{n-1} s_n$
binary word-form features	intial_Upper_Case, all_Upper_Case, no-initial_Capital_Letter(s), all_Lower_Case, contains_(period/number/hyphen ...)
word length	integer
Sentence info	last_word ('.', '?', '!')


Pierre NNP Vinken NNP , , 61 CD years NNS
old JJ , , will ?? join VB the DT board NN
as IN a DT nonexecutive JJ director NN
Nov. NNP 29 CD . .
w-3:years w-2:old
w-1:, w0:will w1:join
w2:the w3:board Swn:.
w-2,-1:old~, w-1,1:,~join
w-1,0:,~will w0,1:will~join
w1,2:join~the
w-1,0,1:,~will~join w-1,1,2:,~join~the
w-2,-1,0:old~,~will w-2,-1,1:old~,~join
w0,1,2:will~join~the
p-3:NNS p-2:JJ p-1:, p-2,-1:JJ~, p-1,1:,~VB_VBP
p1,2:VB_VBP~DT p-2,-1,1:JJ~,~VB_VBP p-1,1,2:,~VB_VBP~DT
k0:MD~NN k1:VB~VBP k2:DT k3:NN s0~MD:1
s0~NN:1 s1~VB:1 s1~VBP:1 s2~DT:1 s3~NN:1

Outline

- Introduction
- **SVMT tool**
  - **SVMT-learner [Training of SVM classifiers]**
  - **SVMT-tagger [POS-tagging of a given input]**
  - **SVMT-evaluator [Study of tagging results]**
  - **SVMT API [Embedded usage of SVMT-tagger]**
- Evaluation
- Conclusions

SVMT-learner

- Options
  - slinding window: length [def: 5] core position [def: 2]
  - feature set [configurable]
  - feature filtering [default: (2 / 100,000)]
  - C parameter tuning (greedy) [default: disabled]
  - SVM model compression [default: disabled]
  - ambiguous/open-class POS lists may be provided if available. [automatically created by default]

SVMT-learner

- dictionary repairing
  - heuristic [by default]
  - using a list of corrections provided

< **the** 50975 6 CD 1 DT 50959 JJ 7 NN 1 NNP 6 VBP 1 >

< **the** 50975 1 DT 50959 >

SVMT-tagger

- Options
  - tagging scheme
    - greedy [default]
    - sentence-level
  - tagging direction
    - left-to-right [default]
    - right-to-left
    - both left-to-right and right-to-left
  - number of tagging passes (1 or 2) [default: 1]
  - backup lexicon

SVMT-evaluator

gold output + SVMT output = report

- brief report
- known vs. unknown tokens
- level of ambiguity
- class of ambiguity
- part-of-speech study

SVMT API

```
my $svmt = SVMTAGGER::SVMT_load(...);
my @tokens = join(/ +/, 'The SVMTool is now being
                         presented to NLP researchers in Lisbon.');
my $input = SVMTAGGER::SVMT_prepare_input(@tokens);
my $output = SVMTAGGER::SVMT_tag($input, $svmt...);
for my $elem (@{$output}) {
    print $elem->get_word."_".$elem->get_pos;
}
```

Outline

- Introduction
- SVMT tool
- **Evaluation**
  - **English on WSJ**
  - **Spanish on LEXESP**
- Conclusions

Evaluation

learning time	1-20 cpu hour
tagging speed	1500 words/second

- 2Ghz Pentium-IV processor; 1Gb RAM
  Perl v5.005_03 (Benchmark package for timing)

Evaluation for English

- Wall Street Journal
  [Penn Treebank III]
- 1,17 million words
  [Training (912k), Validation (132k) and Test (130k)]
- Penn Treebank tagset -> 48 tags
  35 parts-of-speech present ambiguity
  17 are open-classes

	TnT	Collins 02	SVMT	Toutanova et al.
Accuracy	96.46%	97.11%	97.16%	97.24%

Evaluation for Spanish

- LEXESP
- 106k words [Training (86k) and Test (20k)]
- Parole tagset -> 183 tags -> 61 tags (reduced tagset)

  43 parts-of-speech present ambiguity

  12 are open-classes

	TnT	SVMT
Accuracy	96.50%	96.89%

Outline

- Introduction
- SVMT tool
- Evaluation
- **Conclusions**

Conclusions

- **highly accurate:** 97.0 - 97.2% [English on WSJ]
- **efficient:** linear SVMs, primal formulation
- **robust:** soft margin SVMs, two-passes, LR + RL
- **very flexible:** rich feature set, tagging strategies
- **portable:** applied to English, Spanish and Catalan
- **simple:** ease to configure, tune and use

Ongoing Steps

- C++ version coming soon
- Study of more flexible and robust tagging schemes
- Better guessing of unknown words
- Unsupervised learning

Thanks

you may download SVMTool v 1.2 at

http://www.lsi.upc.es/~nlp/SVMTool

References

1. T. Brants. "TnT - A Statistical Part-of-Speech Tagger". In Proceedings of the Sixth ANLP, 2000.
2. T. Nakagawa and T. Kudoh and Y. Matsumoto. "Unknown Word Guessing and Part-of-Speech Tagging Using Support Vector Machines". In Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium, 2001.
3. M. Collins. "Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms", In Proceedings of the 7th EMNLP Conference, 2002.
4. K. Toutanova and D. Klein and C. D. Manning. "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network". In Proceedings of HLT-NAACL'03.
5. J. Giménez and L. Màrquez. "Fast and Accurate Part-of-Speech Tagging: The SVM Approach Revisited". In Proceedings of RANLP '03.
6. N. Cristianini and J. Shawe-Taylor. "An Introduction to Support Vector Machines", Cambridge University Press, 2000.
7. T. Joachims. "Making large-Scale SVM Learning Practical". Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.