Advanced Natural Language Processing and Information Retrieval

LAB2: Kernel Methods for Classification

Alessandro Moschitti

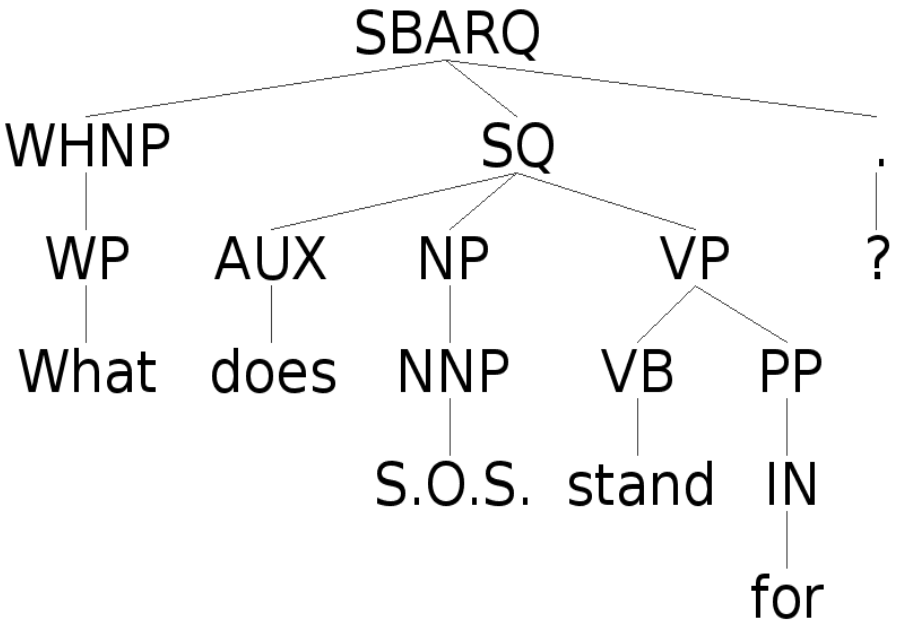
Department of Computer Science and Information Engineering
University of Trento
Email: moschitti@disi.unitn.it

SVM-light-TK Software

- Encodes STK, PTK and combination kernels in SVM-light [Joachims, 1999]
- Available at http://disi.unitn.it/moschitti
- Tree forests, vector sets
- You can download the latest version and other material at the Tutorial Webpage:
  - http://disi.unitn.it/moschitti/SIGIR-tutorial.htm
    - click on SIGIR 2013 Exercise 1

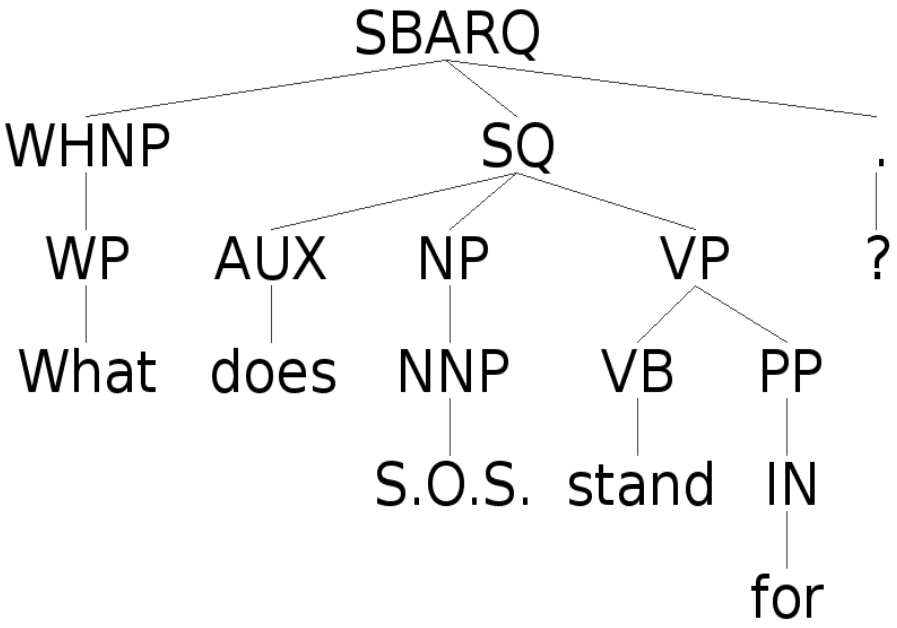
Data Format

- **“What does S.O.S. stand for?”**
- 1 |**BT**| (SBARQ (WHNP (WP What))(SQ (AUX does)(NP (NNP S.O.S.))(VP (VB stand)(PP (IN for))))(. ?))



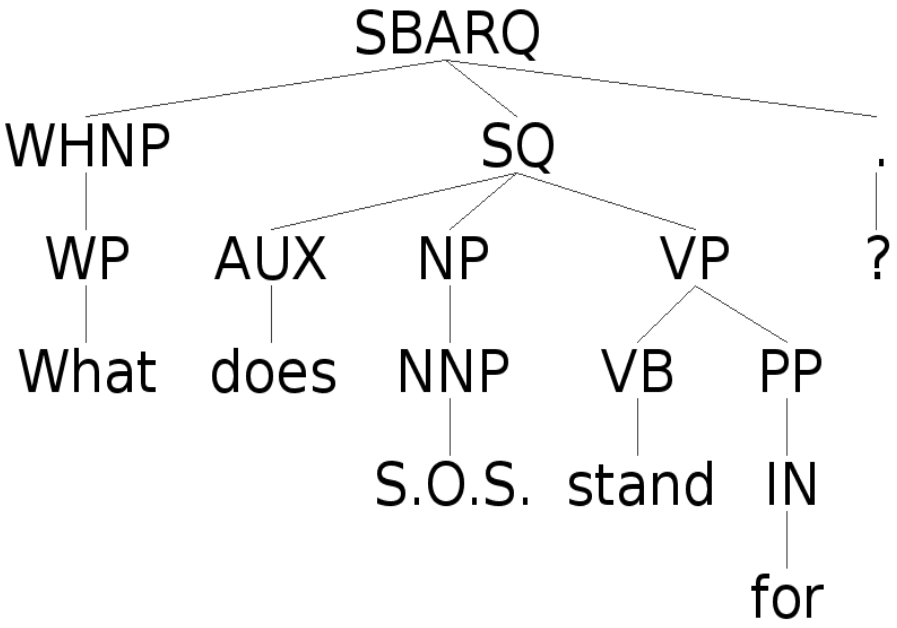
Data Format

- **“What does S.O.S. stand for?”**
- 1 |**BT**| (SBARQ (WHNP (WP What))(SQ (AUX does)(NP (NNP S.O.S.))(VP (VB stand)(PP (IN for))))(. ?))



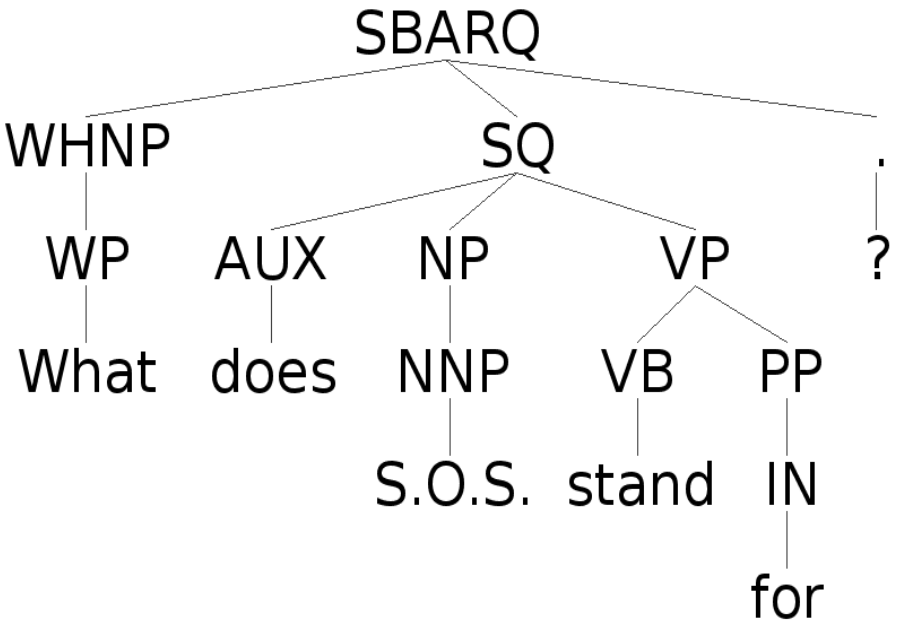
Data Format

- **“What does S.O.S. stand for?”**
- 1 |**BT**| (SBARQ (WHNP (WP What))(SQ (AUX does)(NP (NNP S.O.S.))(VP (VB stand)(PP (IN for))))(. ?))



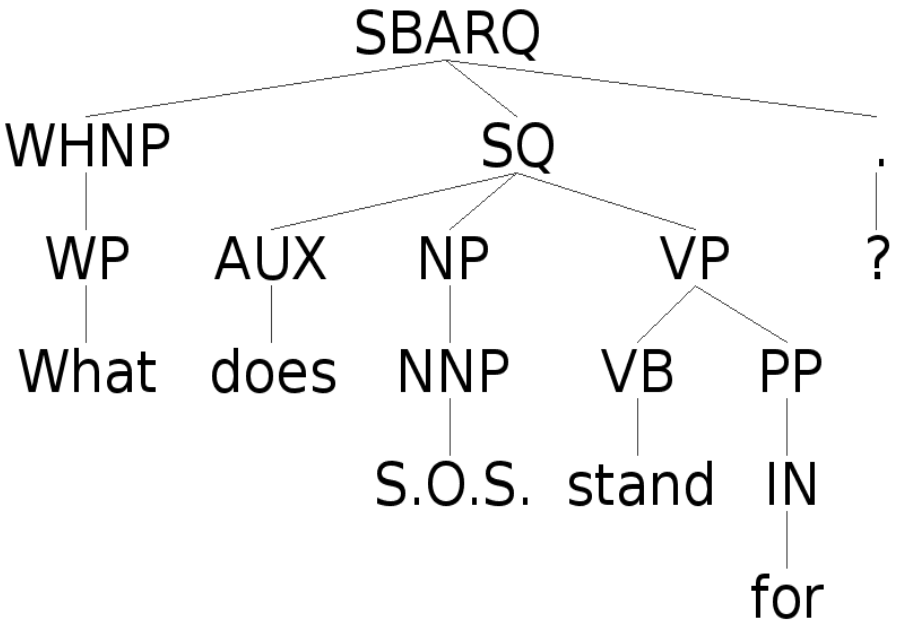
Data Format

- **“What does S.O.S. stand for?”**
- 1 |**BT**| (SBARQ (WHNP (WP What))(SQ (AUX does)(NP (NNP S.O.S.))(VP (VB stand)(PP (IN for))))(. ?))



Data Format

- **“What does S.O.S. stand for?”**
- 1 |**BT**| (SBARQ (WHNP (WP What))(SQ (AUX does)(NP (NNP S.O.S.))(VP (VB stand)(PP (IN for))))(. ?))



Data Format

- **“What does S.O.S. stand for?”**
- 1 **|BT|** (SBARQ (WHNP (WP What))(SQ (AUX does)(NP (NNP S.O.S.))(VP (VB stand)(PP (IN for))))(. ?))

|BT| (***BOW*** (What *)(does *)(S.O.S. *)(stand *)(for *)(? *))

|BT| (***BOP*** (WP *)(AUX *)(NNP *)(VB *)(IN *)(. *))

|BT| (***PAS*** (ARG0 (R-A1 (What *)))(ARG1 (A1 (S.O.S. NNP))) (ARG2 (rel stand)))

|ET| 1:1 21:2.74243946564223 6E-4 23:1 30:1 36:1 39:1 41:1 46:1 49:1 66:1 152:1 274:1 333:1

|BV| 2:1 21:1.44213471486146 54E-4 23:1 31:1 36:1 39:1 41:1 46:1 49:1 52:1 66:1 152:1 246:1 333:1 392:1 **|EV|**

Kernel Combinations an example

K_p^3 : Polynomial kernel of flat features

K_{Tree} : Tree kernel

- Kernel Combinations:

$$K_{Tree+P} = \gamma \times K_{Tree} + K_p^3, \qquad K_{Tree \times P} = K_{Tree} \times K_p^3$$

$$K_{Tree+P} = \gamma \times \frac{K_{Tree}}{\left|K_{Tree}\right|} + \frac{K_p^3}{\left|K_p^3\right|}, \qquad K_{Tree \times P} = \frac{K_{Tree} \times K_p^3}{\left|K_{Tree}\right| \times \left|K_p^3\right|}$$

Basic Commands

- Training and classification
  - **./svm_learn -t 5 -C T train.dat model**
  - **./svm_classify test.dat model**
- Learning with a vector sequence
  - **./svm_learn -t 5 -C V train.dat model**
- Learning with the sum of vector and kernel sequences
  - **./svm_learn -t 5 -C + train.dat model**

LAB2.b: Combining Kernels

- ../SVM-Light-TK-1.5/svm_learn -t 100 -u 1 -j 100 TREC.train model
- ../SVM-Light-TK-1.5/svm_classify TREC.test model

LAB2.b: Combining Kernels

- ../SVM-Light-TK-1.5/svm_learn -t 17 -U 1 -j 300 TREC.train model
- ../SVM-Light-TK-1.5/svm_classify TREC.test model

tree_kernels.param

Type	Type	λ	μ	Norm.	Weight	Comments
1,	1,	.4,	.4,	1,	1	:QUESTION: PT

tree_kernels.param

1,1,.4,.4,1,1 :QUESTION: PT
-10,6,.4,.4,1,1 :BOW
-10,6,.4,.4,1,1 :POS
-10,1,.4,1,1,1 :?
-10,1,.4,1,1,1 :?
-10,1,.4,1,1,1 :?
-10,1,.4,1,1,1 :?
-10,1,.4,1,1,1 :?
-10,1,.4,1,1,1 :?
-10,1,.4,1,1,1 :?
-10,1,.4,1,1,1 :PAS 0
-10,1,.4,1,1,1 :PAS 1
-10,1,.4,1,1,1 :PAS 2
-10,1,.4,1,1,1 :PAS 3
-10,1,.4,1,1,1 :PAS 4 (line 14)

tree_kernels.param

```
1,1,.01,.4,1,1  :ANSWER:PT
-10,6,.4,.4,1,1  :ANSWER:BOW
6,6,.4,1,1,1 :ANSWER:POS
-10,1,.4,1,1,1 :ANSWER:?
-10,1,.4,1,1,1 :ANSWER:?
3,3,.4,.1,1,1   :ANSWER:PAS 0 (line 20)
3,3,.4,.1,1,1   :ANSWER:PAS 1
3,3,.4,.1,1,1   :ANSWER:PAS 2
-10,1,.4,1,1,1 :ANSWER:PAS 3
-10,1,.4,1,1,1 :ANSWER:PAS 4
10000,0,0,0,0,0: END_OF_TREE_KERNELS
```

LAB2.C: Smoothing Partial Tree Kernel

- src/svm_learn -t 6 -F 5 **-H .5** -X 1 -A 0 -j 2
- -C + -W R -V R
- -P LSA250-dim.txt
- qc_coarse_dataset/LCT/ABBR_train.dat
- model

- src/svm_classify qc_coarse_dataset/LCT/ABBR_test.dat model