TILBURG UNIVERSITY

Machine learning for sequential data:

A comparative study with applications to natural language processing

Sander Canisius

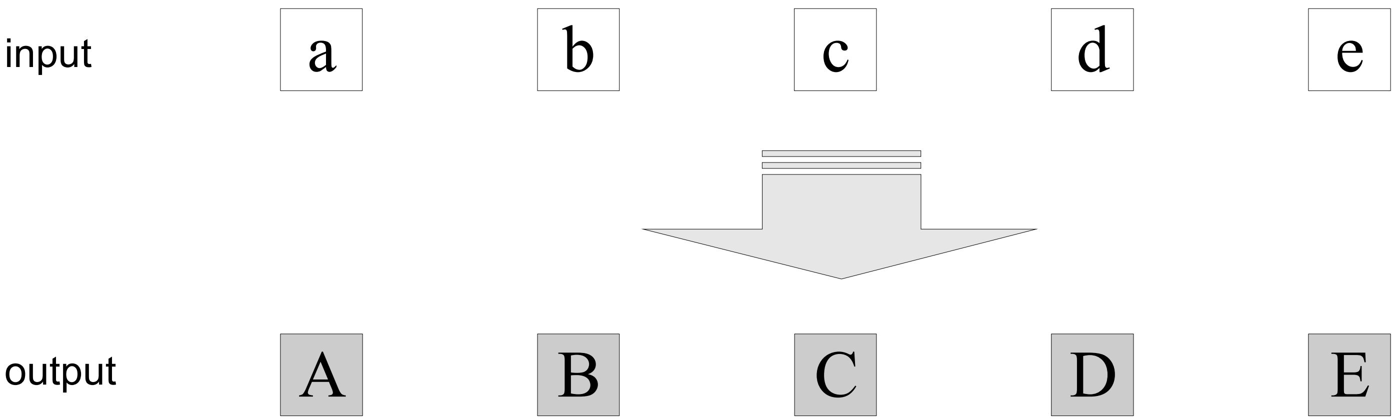
S.V.M.Canisius@uvt.nl

ILK / Language and Information Science

Tilburg University

(Joint work with Antal van den Bosch and Walter Daelemans)

Predicting label sequences



Sequences in NLP

Olympic champion Agassi meets Karim Alami of Morocco in the first round .

```
 JJ         NN        NNP     VBZ     NNP    NNP  IN    NNP  IN  DT  JJ   NN    .
I-NP       I-NP      I-NP     I-VP    I-NP   I-NP I-PP I-NP   I-PP I-NP I-NP I-NP

[NP                     ]   [VP   ]  [NP          ][PP][NP   ][PP][NP            ]

 I-MISC      O        I-PER      O     I-PER I-PER O  I-LOC    O   O    O    O

 [MISC  ]            [PER   ]          [PER          ]   [LOC    ]
```

Sequences in NLP

```
p r e e x i s t i n g
p r i I G I s t I - N
c 0 0 c 0 0 0 0 i 0 0
[c   ][c       ][i   ]
```

Machine learning methods

Conditional Random Fields	(Lafferty et al., 2001)
Hidden Markov SVM, Label Sequence AdaBoost	(Altun & Hofmann, 2003)
Cycling Dependency Networks	(Toutanova et al., 2003)
Max-margin Markov Networks	(Taskar et al., 2003)
Conditional Markov Models	(Ratnaparkhi, 1996)
Maximum-entropy Markov Models	(McCallum et al., 2000)
Discriminatively trained Hidden Markov Models	(Collins, 2002)
Stacked Sequential Learning	(Cohen, 2004)
Constraint Satisfaction Inference	(Canisius et al., 2006)

…

Benchmark data sets

Natural language processing

- Word-level
  - CELEX Morphological segmentation / parsing
  - CELEX Grapheme-phoneme conversion
- Sentence-level
  - CoNLL-2000 Syntactic chunking
  - CoNLL-2002/3 Named-entity recognition
  - GENIA Named-entity recognition
- Document level
  - FAQ segmentation

Benchmark data sets

Bioinformatics

- Protein secondary structure prediction
- Gene prediction

Suggestions?

- ...

TILBURG ◆ UNIVERSITY

Case study: (bio)medical named-entity recognition

Named-entity recognition in Medline abstracts

[DNA_part Ii kappa B-1] is a [DNA positive regulatory element] in [cell_line B-cell lines] and in the [cell_line Ii-expressing T-cell line] , [cell_line H9] , but acts as a [DNA negative regulatory element] in [cell_line myelomonocytic] and [cell_line glia cell lines] .

Case study: (bio)medical named-entity recognition

Named-entity recognition in Dutch medical encyclopedias

[duration Tussen het vierde en tiende jaar] kunnen [symptom vetophopingen] ([symptom xanthoma 's]) in de [body_part huid] ontstaan .

TILBURG ◆ UNIVERSITY

Learning method

- Maximum-entropy models
  - a.k.a log-linear models

$$P(c|d,\lambda)=\frac{\exp(\sum_i \lambda_i f_i(c,d))}{\sum_{c'} \exp(\sum_i \lambda_i f_i(c',d))}$$

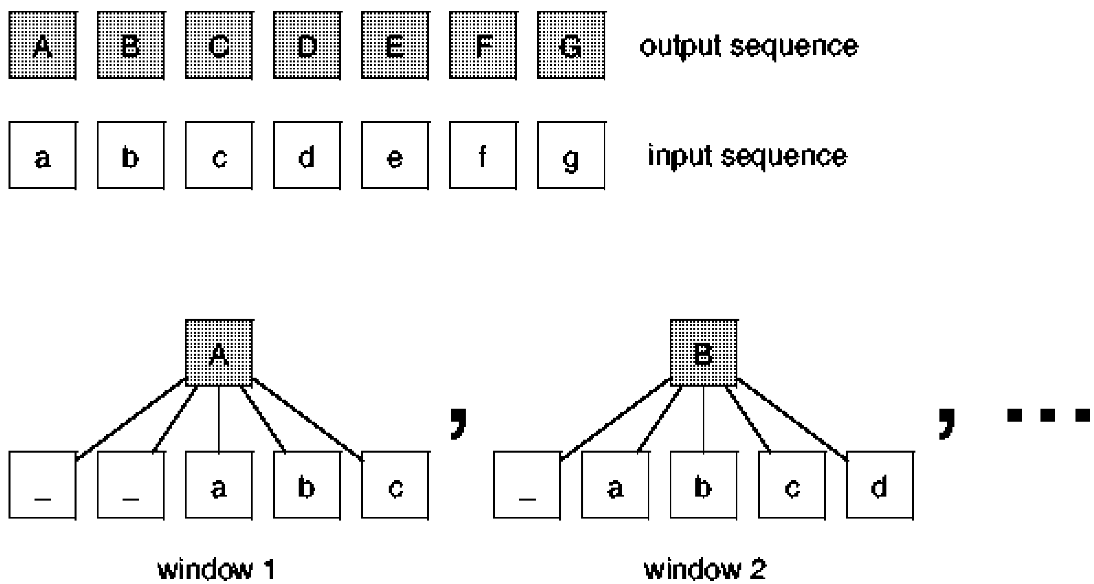
TILBURG ◆ UNIVERSITY

Sequence prediction methods

- Sliding window
- Recurrent sliding window
- Stacking
- Constraint satisfaction inference
- Conditional markov models
- Maximum-entropy markov models
- Conditional random fields

Features

- Simple features only
  - 3-1-3 sliding window of words and POS tags



Results

GENIA

Method	Precision	Recall	$F_{\beta=1}$
Sliding window	54.9 ±1.16	54.1 ±1.23	54.5 ±1.02
Rec. sliding window	67.3 ±1.04	57.6 ±1.25	62.1 ±1.11
Stacking	57.8 ±1.21	55.3 ±1.07	56.5 ±1.11
CSI	64.1 ±1.06	56.6 ±1.10	60.1 ±1.03
CMM	67.7 ±0.96	57.9 ±1.07	62.4 ±1.01
MEMM	67.1 ±1.14	57.7 ±1.13	62.1 ±1.15
CRF	66.8 ±1.10	59.2 ±1.14	62.8 ±1.08

Results

Dutch medical encyclopedia

Method	**Precision**	**Recall**	$F_{\beta=1}$
Sliding window	62.3 ±1.12	60.8 ±1.06	61.5 ±0.98
Rec. sliding window	68.5 ±1.16	60.0 ±1.13	63.9 ±0.89
Stacking	63.2 ±1.23	60.8 ±1.13	62.0 ±1.10
CSI	68.6 ±1.15	59.9 ±1.11	63.9 ±1.02
CMM	68.8 ±1.26	59.6 ±1.09	63.9 ±0.99
MEMM	68.8 ±1.09	59.3 ±1.26	63.7 ±1.09
CRF	66.8 ±1.14	60.2 ±1.14	63.4 ±0.99

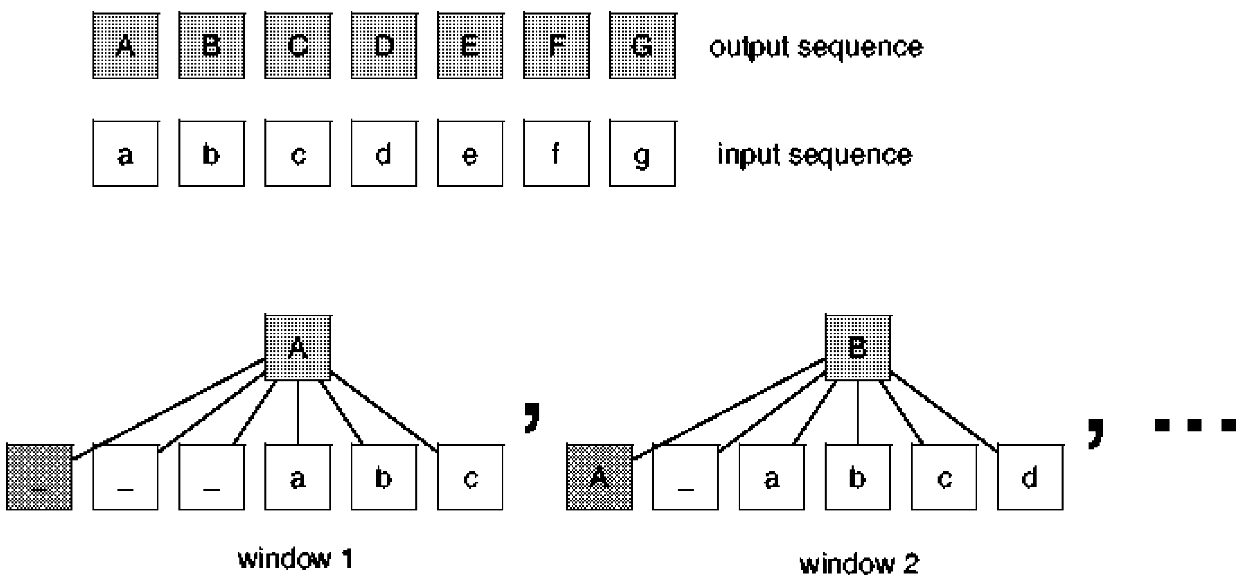
Observations

- Sequence methods tend to favour precision over recall
  - In named-entity recognition tasks, entities are predicted more conservatively

- Very similar performance with many sequence methods

- Recurrent sliding window and its probabilistic version CMM have almost exactly the same performance
  - Doesn't the extra inference step add anything?

Recurrent sliding window



Ratnaparkhi's conditional markov models

- Label sequence conditional probability

$$P(y_{1,}y_{2,}\ldots,y_n|x_{1,}x_{2,}\ldots,x_n)=\sum_i p(y_i|h_i,x_i)$$

- h_i corresponds to the history features in the recurrent sliding window method

- Beam search is used to select to most likely label sequence

FAQ segmentation

McCallum et al., 2000

```
<prolog>          This section of the FAQ is about the electronic support network
<prolog>          that exists for 386bsd and its off-spring.
<prolog>
<question>1.0     I just downloaded all of 386bsd version 0.1 and I can't get
<question>        [some feature] to work?  Do you have any suggestions?
<answer>
<answer>          Yes.  Get FreeBSD, OpenBSD, or NetBSD.
<answer>
<answer>
<question>1.1     Minimum hardware configuration recommended
<answer>
<answer>          There has been considerable debate about what the REAL minimum
<answer>          configuration for *BSD is.  Some would claim that it is the
```

Tilburg University

Features for FAQ segmentation

McCallum et al., 2000

```
begins-with-number
begins-with-ordinal
begins-with-punctuation
begins-with-question-word
begins-with-subject
blank
contains-alphanum
contains-bracketed-number
contains-http
contains-non-space
contains-number
contains-pipe
```

```
contains-question-mark
contains-question-word
ends-with-question-mark
first-alpha-is-capitalized
indented
indented-1-to-4
indented-5-to-10
more-than-one-third-space
only-punctuation
prev-is-blank
prev-begins-with-ordinal
shorter-than-30
```

Results

FAQ segmentation

Method	Precision	Recall	$F_{\beta=1}$
Default maxent	21.3	46.6	27.7
Rec. sliding window	70.8	73.6	70.7
CMM	74.2	78.0	74.9

Discussion

- Recurrent sliding window (CMM, beam size: 1) vs. CMM
  - Hardly any difference on two domain-specific entity recognition tasks
  - CMM outperforms recurrent sliding window on FAQ segmentation
  - What causes these differences?
    - Do the properties that favour CMMs actually occur in real-world NLP tasks?
    - So far, various potential explanations have been explored, none proved to be true

Summary

- Presented plans and preliminary results for a large-scale empirical evaluation of sequence prediction methods in the context of natural language processing
- Suggestions for relevant/informative data sets are welcome
- Small case study on domain-specific entity recognition
  - Sequence prediction methods tend to improve F-score mainly by improving precision, not recall
  - Inference methods on top of probabilistic (maxent) classifiers did not prove to have a large advantage over simpler methods
    - However, there may be data sets where this advantage does exist (e.g. FAQ segmentation)