university of groningen

Structured Prediction for Named Entity Recognition

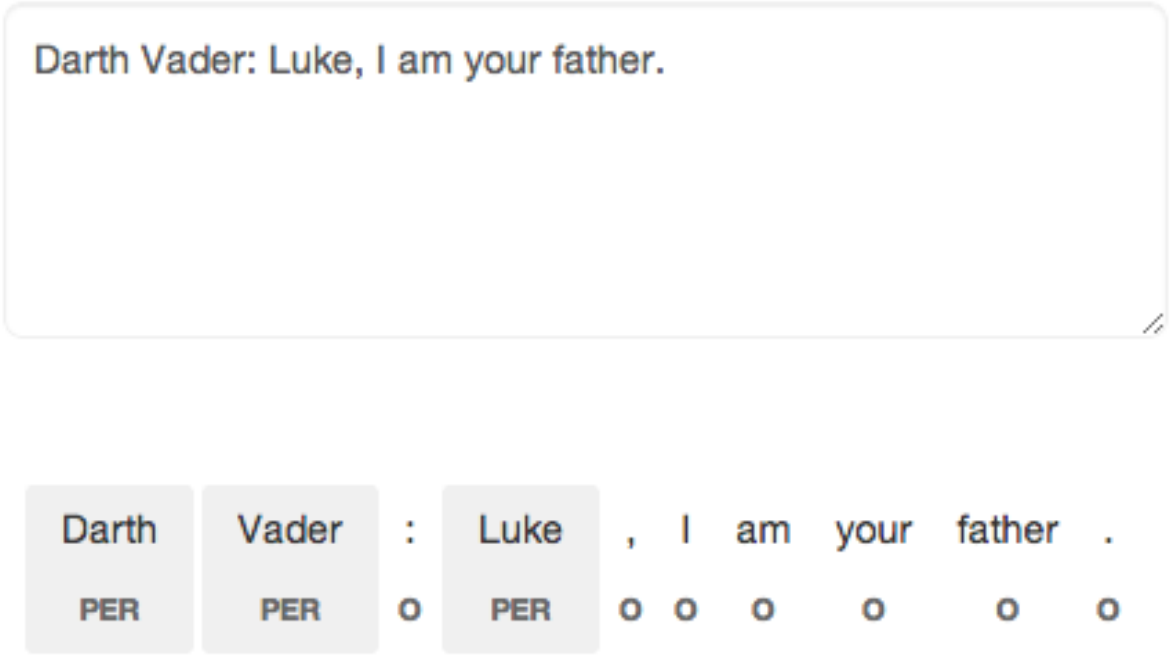
Joachim Daiber & Carmen Klaussner

Information Science
University of Groningen

October 24th, 2012

university of groningen

Named Entity Recognition (NER)



university of groningen

Entity Classes in NER

- Enamex types
  - Person Names: `John Bateman`
  - Organisations: `Lavazza`
  - Locations: `France, Bristol`
- Miscellaneous (CoNLL)
  - proper names outside the classic *enamex*
  - the type *product*
- timex (Date & Time Expressions)
- numex Monetary Values & Percent

$\Rightarrow$ only specific entities; *in June*/ *the prof* (undefined year/person)

university of groningen

Labeled data

- organized challenges
  - CoNLL 2002/2003
  - Languages: English, German, Spanish & Dutch
  - news data
- labels: ORG / LOC/ PER/ MISC

university of groningen

Approaches to NER

1. linguistic grammar-based techniques
2. statistical models

1. hand-crafted rules may obtain a high precision, but at cost of low recall and extensive work by computational linguists
2. Statistical NER systems require large amount of manually annotated training data

$\Rightarrow$ supervised methods most prominent

university of groningen

Issues for NER

- Ambiguity
  - **Polysemy:** `Location` vs. `Person`
    Paris (France) - Paris (Hilton)
  - **Metonymy:** (part-whole):
    "**Paris** has decided to introduce an increase in tax..."
    $\Rightarrow$ (the **government** not the **city**)
- mainly domain-specific systems - not readily portable to different domain/genre

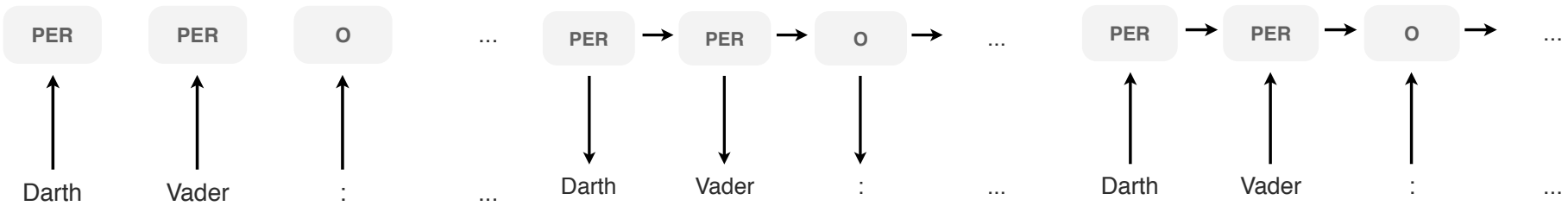
university of groningen

Structured Prediction

- Learning the predictor $\mathbf{x} \rightarrow \mathbf{y}$
- Non-structured prediction: $\mathbf{y}$ is atomic
- Structured prediction: $\mathbf{y}$ is structured (e.g. tree, sequence, etc.)

university of groningen

Structured Prediction



Local classifiers

- features
- no label interactions

HMM

- MLE over tokens
- label interaction

Structured Prediction

- features
- label interactions

university of groningen

Our Implementation

- **Structure:** Labels of the full sentence (O if none)
- **Learning:** Structured Perceptron with Averaging
- **Decoding:** Viterbi algorithm (Markov assumption, only 1 prev. label)

university of groningen

Features

- **Node features**:
  - token, suffixes and prefixes (2-4)
  - number patterns, contains '-', etc.
  - Capitalized?, UPPERCASE?
  - lemma, POS tag
- **Label interaction**:
  - current and prev. label
  - current token and last label for prepositions or possessive 's
- **Gazetteer**:
  - Mark entities from lists of known names. *Reliability* of each list is learnt by the Perceptron.
  - Lists are automatically created using a SPARQL query over DBpedia.

university of groningen

Some Challenges

- **Marking gazetteer entries:**
  - Directed Acyclic Word Graph (Q: Do I know `New York .*` ?)
  - starting at each token, mark longest known entry
- **Headlines and irregular case:**

  `SOCCER - BELARUS BEAT ESTONIA IN WORLD CUP QUALIFIER .`

  - useful case information missing
  - restore most likely case before classification (*truecasing*)

university of groningen

References I


Xavier Carreras.
Learning Structured Predictors.
Lecture at Lisbon Machine Learning School 2012, `http://lxmls.it.pt/strlearn.pdf`, 2012.


Michael Collins.
Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms.
In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 1–8, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

university of groningen

References II


David Nadeau and Satoshi Sekine.
A survey of named entity recognition and classification.
Linguisticae Investigationes, 30(1):3–26, January 2007.
Publisher: John Benjamins Publishing Company.


Erik F. Tjong Kim Sang and Fien De Meulder.
Introduction to the CoNLL-2003 shared task: language-independent named entity recognition.
In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 142–147, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.