Robust Semi-Supervised Named Entity Recognition

Rodrigo Agerri

Outline

1. Introduction
2. Supervised NERC
3. ixa-pipe-nerc
4. Languages
5. Out of Domain Evaluation
6. Conclusion

Outline

1. Introduction
2. Supervised NERC
3. ixa-pipe-nerc
4. Languages
5. Out of Domain Evaluation
6. Conclusion

The disappearance of York University chef Claudia Lawrence is now being treated as suspected murder, North Yorkshire Police said. However detectives said they had not found any proof that the 35-year-old, who went missing on 18 March, was dead. Her father Peter Lawrence made a direct appeal to his daughter to contact him five weeks after she disappeared. His plea came at a news conference held shortly after a 10,000 reward was offered to help find Miss Lawrence. Crimestoppers said the sum they were offering was significantly higher than usual because of public interest in the case.

The disappearance of **York University** chef **Claudia Lawrence** is now being treated as suspected murder, **North Yorkshire Police** said. However detectives said they had not found any proof that the 35-year-old, who went missing on 18 March, was dead. Her father **Peter Lawrence** made a direct appeal to his daughter to contact him five weeks after she disappeared. His plea came at a news conference held shortly after a **10,000** reward was offered to help find **Miss Lawrence.** Crimestoppers said the sum they were offering was significantly higher than usual because of public interest in the case.

Information Extraction

- Find and understand relevant parts of a text.
- Produce an structured view of a text: who did what to whom?
- Structure information for its use in automatic inference.

NERC

- Entities that can be linked, related, indexed, etc.
- Clarify attribution and object of text (target and holder).
- Question Answering systems.
- Browsers, etc.

Outline

1 Introduction

2 Supervised NERC

3 ixa-pipe-nerc

4 Languages

5 Out of Domain Evaluation

6 Conclusion

Supervision

- **Training:**
  - Create a collection of representative documents and label each token with its class.
  - Design feature extractors according to text and classes.
  - Train a sequence labeler to predict the classes from the data.
- **Testing:**
  - A unlabeled set of documents.
  - Label each token with the trained model.

Local Features

- 5 tokens
- Current token
- Shape: Whether it contains numbers, punctuation, etc.
- Previuos decision for current token.
- Beginning of sentence.
- 4 first characters of prefix and suffix.
- Token and their shapes bigrams.

Contextual predicates

- Conditional probability of a history *b* and of a label *a* determined by the parameters whose features are active.
- When a feature is active, its corresponding parameter in the model will contribute to the probability of $p(a|b_i)$.

Conditional Maximum Entropy Models

- Training set: $T = (a_1, b_1) \ldots (a_n, b_n)$ where
  - $(b_1 \ldots b_n)$ is a large set of contexts and
  - $(a_1 \ldots a_n)$ their correct corresponding classes.
- Combinar the features asigning them weights in a discriminative model (condicional).

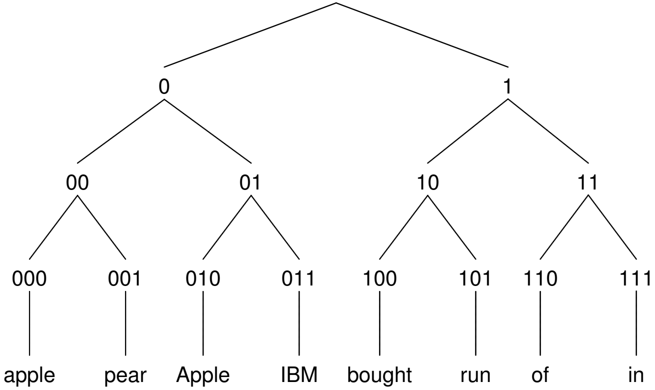
Objectives

- Optimal featureset across languages, domains and datasets.
- On demand and easy generation of NERC systems.
- Keep linguistic annotation to a minimum.
- Reduce (or dispose of) manual feature tuning.
- High performance.

Word class models for NER (Ratinov and Roth 2009)

- Leverage unlabeled text for improving NER. **90.57 CoNLL 2003 results** (Ratinov and Roth 2009).
- Previously used for dependency parsing (Koo et al. 2008), Chinese Word segmentation (Liang et al. 2005), Ando and Zhang (2005), Suzuki and Isozaki (2008).
- Brown hierarchial clustering into a binary tree (Brown 1992).
    - Each word can uniquely identified by its path from the root.
    - The path can be represented with a bit string.
- Liang: `https://github.com/percyliang/brown-cluster`.

Word class models for NER (ii)



- Maximize the mutual information of bigrams.
- We can choose the word class at several levels (4, 8, 12, 20).

Word Representations: Turian et al (2010)

1. Mathematical objects associated with each word.
2. Which word representations are good for which tasks?
3. Should we prefer certain features?
4. Can we combine them?
5. Low real valued embeddings vs word representations in clustering?

Experiments on CoNLL 2003 still showed that Brown clusters performed better than word embeddings.

Skip-gram word2vec

- Passos et al. (2014), very complex system featuring phrase based word embeddings.
  - Stacked (2) linear chain CRFs.
  - Local Features as in (Zhang and Johnson 2003).
  - It includes many manually collected lexicons and gazetteers.
  - “Baseline” 87.93 F1.
  - Lexicon infused phrase-based skip-gram embeddings.
  - F1 90.90 CoNLL 2003 thanks to gazetteers.

Outline

1. Introduction
2. Supervised NERC
3. ixa-pipe-nerc
4. Languages
5. Out of Domain Evaluation
6. Conclusion

Pending questions

- Best systems use word clustering or embeddings.
- Diversity of opinions and results as to which method is better (e.g. works for other languages and domains)
- Combination of clusters mixed results.
- Complex systems (global features, private data, gazetteers, embeddings, linguistic annotations).

Contributions

1. Simple and robust system across languages, datasets and domains.
2. Establish clear guidelines to adapt system to a new domain, language or task.
3. No linguistic annotation (only some supervision to the task).
4. Feature combination of various cluster features (in a window).
    - No manual tuning of features.
    - No manual collection of gazetteers.
    - Unlabelled data.
5. Best results.

Local Features

- 5 token window
- Current token
- Shape: whether the token contains numbers, punctuation, starts with uppercase, etc.
- Previous outcome
- Start of sentence
- Prefix and Suffix (4 characters)
- Bigrams (token and shape)
- Trigrams (token and shape) and Character ngrams (complex morphology).

Simple features based on word clusters

- 5 token window.
- Brown (1992) clusters, 4th, 8th, 12th and 20th nodes.
  - Induce on related corpus for best individual results.
  - Corpus not that large.
  - Not accumulative.
- Clark (2003) clusters.
  - Few clusters with related corpus (not that large).
  - More clusters with large corpus (works with generic corpus too, e.g., Wikipedia).
  - It combines well with other clusters and accumulative across datasets.
- Word2vec (Mikolov et al. 2013) clusters, based on K-means over word embeddings of skip-gram.
  - Requires very large corpora (gigaword, wikipedia) even if out of domain.
  - It provides extra recall.
  - Seriously fast to train.
  - Accumulative across datasets.

ixa-pipe-nerc recipe

1. Local features.
2. Brown 1000 clusters trained with related corpus.
3. Clark $k^3 = n$ where n is the size of the corpus
   - More clusters and larger corpus if unrelated.
   - Accumulative across unlabeled corpora.
4. Word2vec with seriously large unrelated corpus (wikipedia, gigaword).
   - Combines well with other clustering features.
   - Accumulative across unlabeled corpora.

Sequence labelling as external knowledge management!

NERC

NERC	Basque	English	Spanish	Dutch
ixa-pipe-nerc	76.66	91.14	84.16	83.18
Passos et al. 2014	n/a	90.90	n/a	n/a
Ratinov and Roth 2009	n/a	90.57	n/a	n/a
Stanford NER	n/a	88.08	n/a	n/a
CMP (2002-03)	n/a	85.00	81.39	77.05
Eihera	71.31	n/a	n.a	n/a

Outline

1 Introduction

2 Supervised NERC

3 ixa-pipe-nerc

4 Languages

5 Out of Domain Evaluation

6 Conclusion

OTE at ABSA SemEval 2014 and 2015

This place is not good enough, especially the service is disgusting.

System (type)	Precision	Recall	F1 score
Baseline	55.42	43.4	48.68
EliXa (u)	68.93	71.22	**70.05**
NLANGP (u)	70.53	64.02	67.12
EliXa (c)	67.23	66.61	**66.91**
IHS-RD-Belarus (c)	67.58	59.23	63.13

Outline

1 Introduction

2 Supervised NERC

3 ixa-pipe-nerc

4 Languages

5 Out of Domain Evaluation

6 Conclusion

Wikinews

Table: Intra-document Benchmarking with Wikinews.

System	mention extent	Precision	Recall	F1
ixa-pipe-nerc	Inner phrase	62.24	77.54	**69.95**
Stanford NER	Inner phrase	62.82	68.69	65.62
ixa-pipe-nerc	Inner token	71.12	80.04	**75.32**
Stanford NER	Inner token	75.56	71.77	73.62
ixa-pipe-nerc	Outer phrase	51.97	68.46	**59.09**
Stanford NER	Outer phrase	50.82	58.75	54.50
ixa-pipe-nerc	Outer token	72.61	66.82	**69.59**
Stanford NER	Outer token	76.83	59.68	67.18

CoNLL 2003 to MUC7

Features	Precision	Recall	F1
MUC7 local	80.09	70.21	74.83
CoNLL local	78.12	60.57	68.23
MUC7 Brown 1000 Reuters	87.13	82.72	**84.86**
CoNLL Brown 1000 Reuters	83.51	77.00	80.12
MUC7 All	90.44	86.84	88.60
CoNLL All	88.47	80.07	**84.06**

Outline

1. Introduction
2. Supervised NERC
3. ixa-pipe-nerc
4. Languages
5. Out of Domain Evaluation
6. Conclusion

Concluding Remarks

- Best results across multiple tasks, languages and domains.
- Robust and simple featureset.
- Exhaustive comparison of word clustering features.
- Combination and accumulation of clustering features instead of complex feature engineering.
- No linguistic annotation required.
- Future work on SST.
- Domain and language adaptation with bootstrapping, silver standards, automatically created resources.
- Cross fertilization with Apache OpenNLP.
- `ixa2.si.ehu.es/ixa-pipes`