Large-Scale Fine-Grained Named Entity Classification

Anette Frank
Department of Computational Linguistics
Heidelberg University

KONVENS 2010
Sept. 6-8 2010, Saarbrücken
Large-Scale Fine-Grained Named Entity Classification

Can we characterize the meanings of Named Entities and their classes distributionally?

Anette Frank
Department of Computational Linguistics
Heidelberg University

KONVENS 2010
Sept. 6-8 2010, Saarbrücken
Joint work with

Asif Ekbal
Eva Sourjiková
Carina Silberer
Simone Paolo Ponzetto

Named Entity Recognition and Classification (NERC): A solved task?

Coarse-grained (multilingual) NERC established in MUC-7 and CoNLL 2003 shared tasks PERS, LOC, ORG, MISC F1-scores in the high 80s

Useful, but:
Can we assign more fine-grained semantic classes to Named Entities in context?

Useful for any information access task
Information Retrieval
Information Extraction
Question Answering
Named Entity Recognition and Classification (NERC): A solved task?

Coarse-grained (multilingual) NERC established in MUC-7 and CoNLL 2003 shared tasks
PERS, LOC, ORG, MISC
F1-scores in the high 80s

Useful, but:
Can we assign more fine-grained semantic classes to Named Entities in context?

Useful for higher-level NLP tasks
Semantic Role Labeling
Coreference Resolution
Selectional Restrictions
Named Entity Recognition and Classification (NERC): A solved task?

Coarse-grained (multilingual) NERC established in MUC-7 and CoNLL 2003 shared tasks PERS, LOC, ORG, MISC
F1-scores in the high 80s

Useful, but:
Can we assign more fine-grained semantic classes to Named Entities in context?

Topic of this talk
How best to address fine-grained NERC?
How hard is the task if pursued at a large scale?
Large-scale fine-grained NERC

FG-NERC
What is it? Why should we care?
What does it involve?

Assessing the challenge of large-scale FG-NERC
First experiments on large-scale FG-NERC
Open issues and extensions

Distributional vs. knowledge-based accounts
[Named Entities in Context]

Conclusion
A case for corpus-based structured distributional models?
Recognizing fine-grained NE classes

What's in a Name?
(Shakespeare, 1597, Romeo and Juliet, Act II, Scene II)

Named Entities

Entities referred to by proper names
What’s in a name?

Unlike common nouns, *proper nouns do not have inherent descriptive meaning.*

Proper nouns have *reference but not sense.* They identify their referents by making use of an arbitrary association that holds between a proper noun and its referent.

Kripke (1980): *Naming and necessity*  
Proper nouns are *rigid designators.*
Recognizing fine-grained NE classes

What's in a Name? (Shakespeare, 1597)

Ontology seeks to provide a definitive and exhaustive classification of entities in all spheres of being. (Smith, 2003)

Named Entities

Given proper names as rigid designators for entities, find their appropriate entity class as defined in large and fine-grained ontologies.
Recognizing fine-grained NE classes

What's in a Name? (Shakespeare, 1597)

... Ontology... derives from the Greek *onto* (being) and *logia* (written or spoken discourse). (Gord Larose; Price Kruse, 2005)

Named Entities

Given proper names as rigid designators for entities, find their appropriate entity class as defined in large and fine-grained ontologies.
Recognizing fine-grained NE classes

How hard is this task?
Recognizing fine-grained NE classes

Easy...

However, XYZ has been composing and performing for many years both in his native Argentina and in the US.
Recognizing fine-grained NE classes

Easy...

However, *Santaolalla* [composer] has been composing and performing for many years both in his native Argentina and in the US.
Recognizing fine-grained NE classes

Needs more context, but still easy ...

complete the fortifications of Famagusta according to the latest theories of defence. XYZ built the great rampart along the sea from the arsenal to the sea castle, entirely surrounding the latter with the new wall. The rampart was thence extended along the northern side of the city to the great Martinengo bastion at the nort-west corner. The Martinengo bastion, the crowning feature of XYZ’s work, is ...
Recognizing fine-grained NE classes

Needs more context, but still easy ...

complete the fortifications of Famagusta according to the latest theories of defence. Sammichele [engineer] built the great rampart along the sea from the arsenal to the sea castle, entirely surrounding the latter with the new wall. The rampart was thence extended along the northern side of the city to the great Martinengo bastion at the north-west corner. The Martinengo bastion, the crowning feature of Sammichele’s work, is ...

Recognizing fine-grained NE classes

More involved...

*In 1871 she married XYZ (well known afterwards for his long connection with The Times), then a widower with three young children. Mr and Mrs XYZ lived in England for the next six years, and thereafter for 11 years more divided their life between England and Italy, where Mr XYZ was correspondent for The Times at Rome.*
Recognizing fine-grained NE classes

More involved...

In 1871 she married William James Stillman [journalist] (well known afterwards for his long connection with The Times), then a widower with three young children. Mr and Mrs Stillman lived in England for the next six years, and thereafter for 11 years more divided their life between England and Italy, where Mr Stillman was correspondent for The Times at Rome.
Recognizing fine-grained NE classes

Tricky ...

One recurrent element was a ship, whether steamer or rowing boat. I would presume these were linked to the artist’s youth, spent in Odessa on the shores of the Black Sea (otherwise he lived in a landlocked life). It was in Odessa that he had first learnt to play cello and piano. XYZ’s writing on music might have expressed his nostalgic longing for that past, as did the recurrent boats in his paintings.
Recognizing fine-grained NE classes

Tricky ...

One recurrent element was a ship, whether steamer or rowing boat. I would presume these were linked to the artist’s youth, spent in Odessa on the shores of the Black Sea (otherwise he lived in a landlocked life). It was in Odessa that he had first learnt to play cello and piano. Kandinsky [painter]‘s writing on music might have expressed his nostalgic longing for that past, as did the recurrent boats in his paintings.
Recognizing fine-grained NE classes

Hopeless ...

The archers would have made use of the convenient debris. The church has a peculiar 900 year old carving of a pigs face to the east of the sough aisle roof. It represents the patron saint of butchers! The church has a number of bright stain glass memorial windows. They are dedicated to XYZ, Mr Wm Quinton, the Reverend E. Davies, the Reverend Thomas Thorpe and Henry Abel Smith (a prominent family of Wilford bankers). At the bottom of the churchyard overlooking the Trent is a small Gazebo (summerhouse) built in 1757. It was often used to store the bodies of locals drowned in ...
Recognizing fine-grained NE classes

Hopeless ...

The archers would have made use of the convenient debris. The church has a peculiar 900 year old carving of a pigs face to the east of the sough aisle roof. It represents the patron saint of butchers! The church has a number of bright stain glass memorial windows. They are dedicated to Henry Kirke White [poet], Mr Wm Quinton, the Reverend E. Davies, the Reverend Thomas Thorpe and Henry Abel Smith (a prominent family of Wilford bankers). At the bottom of the churchyard overlooking the Trent is a small Gazebo (summerhouse) built in 1757. It was often used to store the bodies of locals drowned in ...
Challenges of FG-NERC

Challenges rise with increasing class inventory

WSD

Fixed set of senses per lexeme to disambiguate

Dictionary size: some hundreds of thousands of words

FG-NE(R)C

Proper names: no inherent descriptive meaning

=> May be assigned any of all known NE classes

Dictionary size: infinite and dynamic

Aims: Assessing the difficulty of FG-NERC
Assessing the challenge of large-scale fine-grained NERC

How to determine NE classes?
How many, how fine-grained?
How difficult is the task?
How best to solve it?

*corpus-based distributional* $\leftrightarrow$ *knowledge-based*
How to determine NE classes?
Prior work on FG-NERC

Early approaches: manual selection of classes

Alfonseca & Manandhar 2002
Concept hierarchy learning & instance classification w/ topic maps
Applied to a handful of core classes
First to relate the problem to WSD

Fleischman & Hovy 2001, 2002
8 high-frequent person and location classes
Bootstrapping training sets from manually tagged seeds
Lexical features and topic signatures
Performance drop on held-out data: 83.5 -> 70%
How to determine NE classes?
Prior work on FG-NERC

Current: reference taxonomies and query logs

Tanev & Magnini (2006)
10 person & location classes from WordNet

People Ontology: WordNet sub-hierarchy of 21 person synsets

NE class acquisition from query logs: 4,583 classes

Bunescu & Paşca (2006)
NE disambiguation based on Wikipedia disambiguation pages
PeopleByOccupation: 110 / 540 / 2,847 classes
How to determine NE classes?

Prior work on FG-NERC

Current: reference taxonomies and query logs

- Tanev & Magnini (2006)
- Pașca (2007), Pașca & van Durme (2009)
- Bunescu & Pașca (2006)

None of these perform recognition in context

Classification only can concentrate on indicative samples

Recognition and classification needs to judge each occurrence
How to determine NE classes?

Exploiting abundance of NEs (and categories) in Wikipedia?

Ponzetto and Strube (2007)
Ponzetto and Navigli (2009)

Deriving a taxonomy from Wikipedia
Linking Wikipedia categories to WordNet synsets

error prone

Suchaneck et al. 2007
Linking WordNet with Wikipedia
How to determine NE classes?

How does *language* mark the semantic class of a named entity?
Not at all, or by *special „introductory constructions“*

**Apposition**
*The earl, who was considered for the film part of the spy James Bond*, lives in a new cottage in Eaton Row ...

**Predicative constructions**
*Dedefptah was a priest of Anubis 2000 years ago and was buried* ..

*Using extraction patterns to acquire a labelled data set for supervised training of distributional semantic models*
Large-scale fine-grained NERC

FG-NERC
What is it? Why should we care?
What does it involve?

Assessing the challenge of large-scale FG-NERC
First experiments on large-scale FG-NERC
Open issues and extensions

Distributional vs. knowledge-based accounts
Named Entities in Context

Conclusion
A case for corpus-based structured distributional models?
Supervised acquisition of a FG-NERC data set

Apposition extraction patterns

(a) [the|The] [JJ|NN]* [NN] [NP]
   the abstract painter Kandinsky
(b) [NP] [,]? [a|an|the]* [JJ|NN]* [NN]
   W. Kandinsky, a Russian-born painter, ..

(a) is more constrained
(b) extracts more noise
Supervised acquisition of a FG-NERC data set

Corpus: ukWaC Web corpus (Baroni et al., 2009)
>2 billion tokens, lemmatised, PoS-tagged

1. Heuristic filtering: frequency threshold, non-words, timex,...
2. Simultaneous extraction: class label + instance + context

<table>
<thead>
<tr>
<th>Class label</th>
<th>Instance</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>philosopher</td>
<td>Karl Popper</td>
<td>. . . science has far more answers than empty theologizing and philosophizing. The philosopher Karl Popper tried to gouge holes in science and ultimately . . .</td>
</tr>
</tbody>
</table>
Supervised acquisition of a FG-NERC data set

Corpus: ukWaC Web corpus (Baroni et al., 2009)
>2 billion tokens, lemmatised, PoS-tagged

1. Heuristic filtering: frequency threshold, non-words, timex,...
2. Simultaneous extraction: class label + instance + context

<table>
<thead>
<tr>
<th>Class label</th>
<th>Instance</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>philosopher</td>
<td>Karl Popper</td>
<td>... science has far more answers than empty theologizing and philosophizing. Karl Popper tried to gouge holes in science and ultimately ...</td>
</tr>
</tbody>
</table>
Mapping to a reference taxonomy (WordNet)

Acquired class labels

WordNet synsets
(with instances as direct hyponyms)

=> Structured FG-NERC Resource
Mapping to person#1 subhierarchy

Coverage and accuracy against WordNet

<table>
<thead>
<tr>
<th></th>
<th>Recall against WN person#1 (C w/ inst)</th>
<th>Precision against WN person#1 (C w/ inst)</th>
<th>Precision (manual eval)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ft = 1</td>
<td>87.9 %</td>
<td>20.1 % *</td>
<td>&gt; 76%</td>
</tr>
<tr>
<td>ft = 40</td>
<td>31.1 %</td>
<td>45.8 %</td>
<td>96.7%</td>
</tr>
</tbody>
</table>

* Remaining 79.9% of class labels: matched to WN person#1 w/o inst
Manual evaluation of 20 classes (x 20 contexts): 76% correct person classes
Mapping to WordNet: Special Cases

Matching all class labels against synsets below person #1
\[ \text{ft} = 40: \text{ 153 class labels} \rightarrow \text{146 WordNet synsets; accuracy: 96.7\%} \]

**Ambiguity (10)**
1. architect, designer
2. couturier, fashion designer, ..., designer => majority instance class
3. interior designer, designer, ..., decorator

**Synonymy (7)**
1. playwright, 2. dramatist => collapsing to one class

**Cross-domain ambiguity (5)**
1. guide; carrier => manually removed

**Incorrect mapping (5)**
1. manager => manually corrected
Unsupervised acquisition of a hierarchical NERC data set (person)

<table>
<thead>
<tr>
<th>Level</th>
<th>#C w/ inst</th>
<th>#C</th>
<th>#inst</th>
<th>#inst /C</th>
<th>% of inst</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>29</td>
<td>8</td>
<td>2,662</td>
<td>332</td>
<td>5.49</td>
</tr>
<tr>
<td>3</td>
<td>57</td>
<td>37</td>
<td>18,229</td>
<td>493</td>
<td>37.58</td>
</tr>
<tr>
<td>4</td>
<td>63</td>
<td>46</td>
<td>18,422</td>
<td>401</td>
<td>37.94</td>
</tr>
<tr>
<td>5</td>
<td>37</td>
<td>30</td>
<td>6,231</td>
<td>208</td>
<td>12.84</td>
</tr>
<tr>
<td>6</td>
<td>18</td>
<td>13</td>
<td>2,366</td>
<td>182</td>
<td>4.88</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>5</td>
<td>423</td>
<td>85</td>
<td>0.87</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>2</td>
<td>179</td>
<td>90</td>
<td>0.36</td>
</tr>
<tr>
<td>all</td>
<td>213</td>
<td>141</td>
<td>48,512</td>
<td>344</td>
<td>100</td>
</tr>
</tbody>
</table>

Type-token ratio: 1.52
Manual validation (38 classes x 40 contexts): Accuracy: 96.58%
Unsupervised acquisition of a hierarchical NERC data set (person)

<table>
<thead>
<tr>
<th>Level</th>
<th>#C</th>
<th>#C w/inst</th>
<th>#inst</th>
<th>% of inst</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>29</td>
<td>8</td>
<td>2,662</td>
<td>332</td>
</tr>
<tr>
<td>3</td>
<td>57</td>
<td>37</td>
<td>18,229</td>
<td>493</td>
</tr>
<tr>
<td>4</td>
<td>63</td>
<td>46</td>
<td>18,422</td>
<td>401</td>
</tr>
<tr>
<td>5</td>
<td>37</td>
<td>30</td>
<td>6,231</td>
<td>208</td>
</tr>
<tr>
<td>6</td>
<td>18</td>
<td>13</td>
<td>2,366</td>
<td>182</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>5</td>
<td>423</td>
<td>85</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>2</td>
<td>179</td>
<td>90</td>
</tr>
<tr>
<td>all</td>
<td>213</td>
<td>141</td>
<td>48,512</td>
<td>344</td>
</tr>
</tbody>
</table>

Manual validation (38 classes x 40 contexts): **Accuracy: 96.58%**
How many and how fine-grained classes?

*Level-wise evaluation* of a baseline NERC system, performing *recognition and classification*

=> Adaptating NERC to a hierarchical classification problem

**Semantic classification**

Experiment with *standard NERC techniques* vs. *global (semantic) context information*

Distributional vs. knowledge-based methods?
Binary classification for FG-NERC

Model: Maximum Entropy (MaxEnt)

Binary classifier (one vs. rest) for NE classes in hierarchy

• Propagation of instances from lower to higher levels
• Population of nodes w/o instances
• Other instances (not in subtree) labeled as Outside (O-) a NE class
• Single decision from each binary classifier using MaxEnt confidence
From binary to multiclass classification

**Winner-takes-all (WTA) strategy**: Weighted voting
Tie-breaking: Random selection of output label

**Level-wise classification**
- Combine outputs of *all binary classifiers at the same level of embedding*
- Weighted voting of binary classifiers at any level
- **Weight**: Confidence value (conditional probability of MaxEnt) for the class

**Global classification**
- Combine outputs of *all binary classifiers of the entire taxonomy*
- **Weight**: Product of confidence value and depth of embedding in the taxonomy
## MaxEnt Models: Feature sets

**MaxEnt-A**  
Standard feature set for NER  
Local classification context

- Context words  
- Word prefix and suffix  
- Infrequent word  
- Part-of-Speech (PoS) and chunk information  
- Capitalization  
- Word length  
- Digit and symbol features  
- Dynamic feature

**MaxEnt-B**  
Inspired from WSD*  
Global classification context

- PoS context: \( \text{pos}_{i-3} \ldots \text{pos}_{i+3} \)  
- Local collocation: three words left and right of target word  
- Surrounding content words: 10 most frequent content words in 10-word window

* Lee & Ng (2002)
Benchmarking on coarse-grained NERC

MaxEnt-A feature set

Datasets: CoNLL-2003 shared task

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>PER</td>
<td>83.02</td>
<td>81.40</td>
<td>82.21</td>
</tr>
<tr>
<td>LOC</td>
<td>88.47</td>
<td>88.19</td>
<td>88.23</td>
</tr>
<tr>
<td>ORG</td>
<td>77.20</td>
<td>68.03</td>
<td>72.23</td>
</tr>
<tr>
<td>MISC</td>
<td>81.20</td>
<td>83.92</td>
<td>82.54</td>
</tr>
<tr>
<td>Overall</td>
<td>81.11</td>
<td>80.47</td>
<td>81.77</td>
</tr>
</tbody>
</table>

2 points off F1 of the most similar system (Bender et al. 2003)
7 points below best-performing system (Florian et al. 2003)
no gazetteer, simple feature set, no system combination
Experiments on FG-NERC

Level-wise classification (at increasingly deeper levels)
Global classification

Models and feature sets
MaxEnt-A: Standard NERC feature set
MaxEnt-B: Semantic feature set
MaxEnt-A&B: Joint feature set

Evaluation
Level-wise
Global: exact match vs. hierarchical measure
(Melamed and Resnik, 2000)
Data sets for FG-NERC

Automatically labelled data
extracted w/ appositional patterns (class labels stripped)
60 tokens left/right context each target NE
Manually validated test set
38 classes x 40 contexts (incl. 18 classes of Giuliano 2009)

<table>
<thead>
<tr>
<th>Set</th>
<th># tokens</th>
<th># NEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training (80%)</td>
<td>2,431,041</td>
<td>38,810</td>
</tr>
<tr>
<td>Development (20%)</td>
<td>478,871</td>
<td>9,702</td>
</tr>
<tr>
<td>Test (validated)</td>
<td>181,490</td>
<td>1,520</td>
</tr>
</tbody>
</table>
Balancing data sets

Unbalanced training set
negative:positive examples at topmost level: 63:1
increase of negative (-O) NE classes at with increasing levels

Sampling by factors to 1:5 (positive vs. negative)
factors \( f \) of oversampling from positive set \( P \) to \( P' \)

<table>
<thead>
<tr>
<th>factor ( f )</th>
<th>size of ( P )</th>
<th>min ( P' )</th>
<th>max ( P' )</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 x ( P )</td>
<td>1-2K</td>
<td>20</td>
<td>40K</td>
</tr>
<tr>
<td>15 x ( P )</td>
<td>2K – 5K</td>
<td>30K</td>
<td>75K</td>
</tr>
<tr>
<td>10 x ( P )</td>
<td>5K – 10K</td>
<td>50K</td>
<td>100K</td>
</tr>
<tr>
<td>5 x ( P )</td>
<td>10K – 20K</td>
<td>50K</td>
<td>100K</td>
</tr>
<tr>
<td>2 x ( P )</td>
<td>20K – 50K</td>
<td>40K</td>
<td>100K</td>
</tr>
<tr>
<td>( P )</td>
<td>50K – ...</td>
<td>50K</td>
<td>&gt; 50K</td>
</tr>
</tbody>
</table>
# Level-wise & global NE recognition and classification

| L | #C | #inst/C | MaxEnt-A R | MaxEnt-B R | MaxEnt-A&B R |
|---|----|---------|
|   |    |         | R | P | F1 | R | P | F1 | R | P | F1 |
| 1 | 1  | -       | 98.7 | 85.0 | 91.4 | 95.1 | 83.0 | 88.6 | 99.8 | 87.8 | 93.4 |
| 2 | 29 | 332     | 96.0 | 65.5 | 77.9 | 48.1 | 46.3 | 47.2 | 98.5 | 67.6 | 80.1 |
| 3 | 57 | 493     | 95.3 | 54.3 | 69.3 | 43.3 | 41.1 | 42.2 | 97.6 | 56.3 | 71.4 |
| 4 | 63 | 401     | 86.8 | 52.8 | 65.6 | 41.1 | 37.2 | 39.1 | 88.9 | 54.5 | 67.6 |
| 5 | 37 | 208     | 90.4 | 45.9 | 60.9 | 49.2 | 21.5 | 29.9 | 91.5 | 47.8 | 62.9 |
| 6 | 18 | 182     | 91.6 | 36.9 | 52.6 | 51.7 | 13.2 | 21.1 | 93.5 | 38.3 | 54.4 |
| 7 | 6  | 85      | 89.5 | 31.8 | 46.9 | 42.2 | 10.2 | 16.4 | 91.3 | 33.5 | 49.1 |
| 8 | 2  | 90      | 100 | 19.9 | 66.7 | 87.1 | 8.1 | 14.7 | 100 | 20.2 | 33.6 |
| Global: exact match | | | 85.1 | 43.2 | 57.3 | 61.9 | 26.6 | 37.2 | 88.1 | 45.4 | 59.9 |
| hierarchical | | | 87.7 | 44.8 | 59.4 | 64.5 | 29.5 | 40.5 | 89.9 | 46.3 | 61.1 |
Level-wise & global
NE classification (NEC)

<table>
<thead>
<tr>
<th>L</th>
<th>#C</th>
<th>#inst /C</th>
<th>MaxEnt-A</th>
<th>MaxEnt-B</th>
<th>MaxEnt-A&amp;B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>R</td>
<td>P</td>
<td>F1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>-</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>29</td>
<td>332</td>
<td>85.8</td>
<td>88.6</td>
<td>87.0</td>
</tr>
<tr>
<td>3</td>
<td>57</td>
<td>493</td>
<td>83.9</td>
<td>88.1</td>
<td>85.9</td>
</tr>
<tr>
<td>4</td>
<td>63</td>
<td>401</td>
<td>74.6</td>
<td>85.0</td>
<td>79.5</td>
</tr>
<tr>
<td>5</td>
<td>37</td>
<td>208</td>
<td>78.8</td>
<td>83.4</td>
<td>80.9</td>
</tr>
<tr>
<td>6</td>
<td>18</td>
<td>182</td>
<td>88.5</td>
<td>73.6</td>
<td>80.4</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>85</td>
<td>79.2</td>
<td>76.5</td>
<td>77.8</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>90</td>
<td>82.8</td>
<td>73.8</td>
<td>78.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>81.1</td>
<td>84.2</td>
<td>82.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>83.5</td>
<td>86.2</td>
<td>84.8</td>
</tr>
</tbody>
</table>
## Comparison to baseline

(most frequent class)

<table>
<thead>
<tr>
<th>Level</th>
<th>Baseline</th>
<th>MaxEnt-A</th>
<th>MaxEnt-B</th>
<th>MaxEnt-A&amp;B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F1</td>
<td>R</td>
</tr>
<tr>
<td>1</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>28.4</td>
<td>25.9</td>
<td>27.1</td>
<td>85.8</td>
</tr>
<tr>
<td>3</td>
<td>27.9</td>
<td>23.1</td>
<td>25.1</td>
<td>83.9</td>
</tr>
<tr>
<td>4</td>
<td>18.8</td>
<td>20.4</td>
<td>19.5</td>
<td>74.6</td>
</tr>
<tr>
<td>5</td>
<td>25.8</td>
<td>19.0</td>
<td>21.9</td>
<td>78.8</td>
</tr>
<tr>
<td>6</td>
<td>24.7</td>
<td>7.8</td>
<td>11.9</td>
<td>88.5</td>
</tr>
<tr>
<td>7</td>
<td>19.1</td>
<td>5.3</td>
<td>8.3</td>
<td>79.2</td>
</tr>
<tr>
<td>8</td>
<td>34.2</td>
<td>2.9</td>
<td>5.5</td>
<td>82.8</td>
</tr>
<tr>
<td>global</td>
<td>34.6</td>
<td>18.5</td>
<td>24.1</td>
<td>81.1</td>
</tr>
<tr>
<td></td>
<td>33.0</td>
<td>21.2</td>
<td>25.9</td>
<td>83.5</td>
</tr>
</tbody>
</table>
A caveat

Our training set consists of appositional contexts (with class labels removed)

„Introductory“ contexts may be more indicative of target classes
So, does the models trained on these data extend to realistic data?

„Follow-up“ contexts
Test performance on extended contexts, to ensure realistic setting both for training and testing
The clothing producer Gap has been served a lawsuit by a child over its "back to school" range of clothes. Douglas Ramsbottom, aged 16, is suing the company over their misleading adverts, where children back at school are shown to be having a good time, which he states "is clearly not the case, dawg".

After watching the most recent Gap advert, which depicts young children twirling around, jumping up and down, and generally enjoying themselves to a funky beat in the background, Douglas then went to Bootle Secondary School for Boys, full of high hopes for much new-style free soul dancing, community spirit, and trendy clothing in the new year....
Follow-up contexts

Measuring automatic annotation quality for extracting contexts for follow-up mentions of NEs – following the *one-sense-per-discourse* assumption –

Annotation set: 112 classes (18.5 contexts/class)
Pairs: Introductory + follow-up contexts

Accuracy of automatically acquired training data
Follow-up contexts: 81.24%
Introductory contexts: 96.58%
More solid evaluation

Training and testing on combinations of mixed appositional and follow-up contexts
FG-NERC trained on follow-up contexts (only)

<table>
<thead>
<tr>
<th>Level</th>
<th>Baseline</th>
<th>MaxEnt-A&amp;B Introductory</th>
<th>MaxEnt-A&amp;B Follow-up</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F1</td>
</tr>
<tr>
<td>1</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>28.4</td>
<td>25.9</td>
<td>27.1</td>
</tr>
<tr>
<td>3</td>
<td>27.9</td>
<td>23.1</td>
<td>25.1</td>
</tr>
<tr>
<td>4</td>
<td>18.8</td>
<td>20.4</td>
<td>19.5</td>
</tr>
<tr>
<td>5</td>
<td>25.8</td>
<td>19.0</td>
<td>21.9</td>
</tr>
<tr>
<td>6</td>
<td>24.7</td>
<td>7.8</td>
<td>11.9</td>
</tr>
<tr>
<td>7</td>
<td>19.1</td>
<td>5.3</td>
<td>8.3</td>
</tr>
<tr>
<td>8</td>
<td>34.2</td>
<td>2.9</td>
<td>5.5</td>
</tr>
<tr>
<td>global</td>
<td>34.6</td>
<td>18.5</td>
<td>24.1</td>
</tr>
<tr>
<td></td>
<td>33.0</td>
<td>21.2</td>
<td>25.9</td>
</tr>
</tbody>
</table>
More solid evaluation

Training and testing on combinations of *mixed appositional and follow-up contexts*

... *in the making*
Discussion

Standard NERC feature set plus lexical semantic features yields best results, with balanced precision and recall values. F1-score of 84.9 (86.4) for *global classification* on a large NE class hierarchy.

**Testing performance on realistic data**

Introductory contexts w/ class labels + follow-up contexts
Enhancements

Global classification models
Classification at *lower levels* suffers *from small training set sizes*
Using a global classification model, where classes inherit information from higher classes, could *alleviate sparsity problems of specific classes*

Recognition
Performance for *recognition and classification* degrades by *25 points* to 59.9 (61.1) F1-score
with serious drops at lower levels
=> decouple recognition from classification, with recognition models trained at higher levels
Open issues and extensions

Range of classes and domains
   Extend to full range of semantic sub-domains
   Extend to more complex classes (compound class labels)

Ambiguity
   *Cross-domain* class label ambiguity
   Entity ambiguity: *multiple roles* vs. *distinct identity*
   (→ Web People Search)

Global classification decisions
   Hierarchical models
   Integrate classifications for entity chains
Open issues and extensions

Range of classes and domains
Extend to full range of semantic sub-domains
Extend to more complex classes (compound class labels)

Ambiguity

Cross-domain class label ambiguity
Entity ambiguity: multiple roles vs. distinct identity
(→ Web People Search)

Global classification decisions
 Hierarchical models
 Integrate classifications for entity chains
Large-scale fine-grained NERC

FG-NERC
What is it? Why should we care?
What does it involve?

Assessing the challenge of large-scale FG-NERC
First experiments on large-scale FG-NERC
Open issues and extensions

Distributional vs. knowledge-based accounts
[Named Entities in Context]

Conclusion
A case for corpus-based structured distributional models?
How best to approach FG-NERC?

Distributional vs. knowledge-based accounts

Relating FG-NERC to WSD
WSD: distributional vs. knowledge-based techniques
Approaching FG-NERC from a knowledge-based WSD perspective

Similar to WSD, an alternative to a supervised setting is to exploit the knowledge contained in a lexical knowledge base (here WordNet)

FG-NEC using Extended Lesk
(Banerjee & Pedersen, 2003)

FG-NEC using Personalised Page Rank
(UKB, Agirre & Soroa, 2009)

Both take into account the context of the target word and take advantage of the connectivity in the semantic network
FG-NERC using kb-WSD methods

Experiment I: Extended Lesk
Performing *level-wise* and *global FG-NEC* on previous data set considering NEs as unknown lexical items possible senses: all entity classes

**Extended Lesk** (Banerjee & Pedersen, 2003)
Computes highest gloss overlap with glosses of all senses of the context words, and those of related concepts (hypernyms, ...
## Extended Lesk vs. MaxEnt-A&B (NEC)

<table>
<thead>
<tr>
<th>L</th>
<th>#C</th>
<th>#inst/C</th>
<th>Baseline R</th>
<th>Baseline P</th>
<th>Baseline F1</th>
<th>Extended Lesk R</th>
<th>Extended Lesk P</th>
<th>Extended Lesk F1</th>
<th>MaxEnt-A&amp;B R</th>
<th>MaxEnt-A&amp;B P</th>
<th>MaxEnt-A&amp;B F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>-</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>29</td>
<td>332</td>
<td>28.4</td>
<td>25.9</td>
<td>27.1</td>
<td>16.4</td>
<td>19.7</td>
<td>17.9</td>
<td>87.7</td>
<td>90.8</td>
<td>89.2</td>
</tr>
<tr>
<td>3</td>
<td>57</td>
<td>493</td>
<td>27.9</td>
<td>23.1</td>
<td>25.1</td>
<td>16.2</td>
<td>16.2</td>
<td>16.2</td>
<td>86.4</td>
<td>90.2</td>
<td>97.8</td>
</tr>
<tr>
<td>4</td>
<td>63</td>
<td>401</td>
<td>18.8</td>
<td>20.4</td>
<td>19.5</td>
<td>11.3</td>
<td>11.3</td>
<td>11.3</td>
<td>76.9</td>
<td>87.8</td>
<td>81.9</td>
</tr>
<tr>
<td>5</td>
<td>37</td>
<td>208</td>
<td>25.8</td>
<td>19.0</td>
<td>21.9</td>
<td>13.5</td>
<td>14.0</td>
<td>13.8</td>
<td>81.4</td>
<td>85.6</td>
<td>83.7</td>
</tr>
<tr>
<td>6</td>
<td>18</td>
<td>182</td>
<td>24.7</td>
<td>7.8</td>
<td>11.9</td>
<td>33.2</td>
<td>37.5</td>
<td>35.2</td>
<td>90.2</td>
<td>75.9</td>
<td>83.4</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>85</td>
<td>19.1</td>
<td>5.3</td>
<td>8.3</td>
<td>49.4</td>
<td>49.4</td>
<td>49.4</td>
<td>81.4</td>
<td>78.9</td>
<td>80.1</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>90</td>
<td>34.2</td>
<td>2.9</td>
<td>5.5</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>84.9</td>
<td>75.3</td>
<td>79.8</td>
</tr>
<tr>
<td>all</td>
<td>213</td>
<td>344</td>
<td>34.6</td>
<td>18.5</td>
<td>24.1</td>
<td>36.5</td>
<td>38.6</td>
<td>37.5</td>
<td>83.4</td>
<td>86.5</td>
<td>84.9</td>
</tr>
</tbody>
</table>

Hardly exceeding the baseline!
FG-NERC using kb-WSD methods

Experiment II: UKB
Performing *global FG-NEC on all vs. ambiguous NEs*
(a) Treating all NEs as unknown entities
(b) Performing disambiguation on known NEs only
(with different roles or identity)

**Personalised Page Rank** (Agirre & Soroa, 2009)
Computing contextualized Page Rank
to determine most prominent sense
FG-NERC using kb-WSD methods

Experiment II: UKB

(a) Global FG-NEC

(b) FG-NED for ambiguous NEs

Disambiguation into **full inventory of person NE classes**

Disambiguation only **for ambiguous (known) NEs**
Ambiguity rate of NEs within person domain

7.2 % of all NEs (types) are ambiguous
94.4% of the ambiguous NEs have 2-3 readings (classes)

<table>
<thead>
<tr>
<th>Classes/NE</th>
<th>#NE</th>
<th>#Inst</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambiguous</td>
<td>1889</td>
<td>9506</td>
</tr>
<tr>
<td>Non-ambiguous</td>
<td>26031</td>
<td>37251</td>
</tr>
<tr>
<td>Overall</td>
<td>27920</td>
<td>46757</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Distribution of ambiguous classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classes/NE</td>
</tr>
<tr>
<td>------------</td>
</tr>
<tr>
<td>23</td>
</tr>
<tr>
<td>16</td>
</tr>
<tr>
<td>14</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>1889</td>
</tr>
</tbody>
</table>
FG-NEC using UKB: Results

<table>
<thead>
<tr>
<th>(a) Disambiguation into full set of classes (w/o NEs in context)</th>
<th>(b) Classifying ambiguous NEs only (w/o NEs in context)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#Instances</td>
</tr>
<tr>
<td>All</td>
<td>46,757</td>
</tr>
<tr>
<td>Ambiguous only</td>
<td>9,506</td>
</tr>
</tbody>
</table>

Global FG-NEC: Devastating!

Still lagging behind classical WSD performance (60-70% for nouns)
FG-NEC using UKB: Results

### (a) Disambiguation into full set of classes (w/o NEs in context)

<table>
<thead>
<tr>
<th></th>
<th>#Instances</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>46,757</td>
<td>9.12%</td>
</tr>
<tr>
<td>Ambiguous only</td>
<td>9,506</td>
<td>8.82%</td>
</tr>
</tbody>
</table>

### (b) Classifying ambiguous NEs only (w/o NEs in context)

<table>
<thead>
<tr>
<th></th>
<th>#Instances</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambiguous</td>
<td>9,506</td>
<td>46.51%</td>
</tr>
</tbody>
</table>

### (a) Disambiguation into full set of classes (w/ known NEs in context)

<table>
<thead>
<tr>
<th></th>
<th>#Instances</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>46,757</td>
<td>4.23%</td>
</tr>
<tr>
<td>Ambiguous only</td>
<td>9,514</td>
<td>4.65%</td>
</tr>
</tbody>
</table>

### (b) Classifying ambiguous NEs only (w/ known NEs in context)

<table>
<thead>
<tr>
<th></th>
<th>#Instances</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambiguous</td>
<td>9,514</td>
<td>39.97%</td>
</tr>
</tbody>
</table>

Global FG-NEC: Devastating!

Still lagging behind classical WSD performance (60-70% for nouns)
### Distribution of ambiguous classes

<table>
<thead>
<tr>
<th>Classes/NE</th>
<th>#NE</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>1</td>
<td>0.05%</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>0.05%</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>0.05%</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0.05%</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>0.11%</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>0.16%</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0.05%</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>0.16%</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>0.21%</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>0.42%</td>
</tr>
<tr>
<td>5</td>
<td>27</td>
<td>1.42%</td>
</tr>
<tr>
<td>4</td>
<td>54</td>
<td>2.84%</td>
</tr>
<tr>
<td>3</td>
<td>233</td>
<td>12.24%</td>
</tr>
<tr>
<td>2</td>
<td>1550</td>
<td>81.41%</td>
</tr>
<tr>
<td>1889</td>
<td></td>
<td>100.00%</td>
</tr>
</tbody>
</table>

### Classification results (w/o NEs in context)

<table>
<thead>
<tr>
<th>#Instances</th>
<th>#Correct</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>66</td>
<td>8</td>
<td>12.12</td>
</tr>
<tr>
<td>26</td>
<td>8</td>
<td>30.77</td>
</tr>
<tr>
<td>23</td>
<td>7</td>
<td>30.43</td>
</tr>
<tr>
<td>47</td>
<td>14</td>
<td>29.79</td>
</tr>
<tr>
<td>26</td>
<td>7</td>
<td>26.92</td>
</tr>
<tr>
<td>50</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>18</td>
<td>8</td>
<td>44.44</td>
</tr>
<tr>
<td>39</td>
<td>9</td>
<td>23.08</td>
</tr>
<tr>
<td>75</td>
<td>20</td>
<td>26.67</td>
</tr>
<tr>
<td>163</td>
<td>50</td>
<td>30.67</td>
</tr>
<tr>
<td>327</td>
<td>110</td>
<td>33.64</td>
</tr>
<tr>
<td>447</td>
<td>119</td>
<td>26.62</td>
</tr>
<tr>
<td>1614</td>
<td>603</td>
<td>37.52</td>
</tr>
<tr>
<td>6585</td>
<td>3438</td>
<td>52.15</td>
</tr>
<tr>
<td>9506</td>
<td>4421</td>
<td>46.51%</td>
</tr>
</tbody>
</table>
FG-NERC using knowledge-based WSD?

This simply doesn’t look promising.

WordNet is not a full-fledged knowledge base
Connectivity mostly confined to hierarchical relations

Named Entities and their classification has to do with
*typical actions in which entities are involved*

Composers *compose*. Authors *publish*.
New cameras *take pictures and are presented at the IFA*.
Cars can be *called back to their constructors*. 
FG-NERC using knowledge-based WSD?

This simply doesn’t look promising.

WordNet is not a full-fledged knowledge base
Connectivity mostly confined to hierarchical relations

Named Entities and their classification has to do with
typical actions in which entities are involved

WordNet does not represent the diverse types of
(world) knowledge we have about typical properties,
actions and circumstances that characterize entity classes!
FG-NERC using knowledge-based WSD?

This simply doesn’t look promising.

WordNet is not a full-fledged knowledge base
Connectivity mostly confined to hierarchical relations

Named Entities and their classification has to do with
*typical actions in which entities are involved*

*To a certain extent, WordNet glosses do, and it might be for that reason that Lesk does still better than UKB*
What to conclude?

1. If WSD is a hard problem ...  
   *fine-grained Named Entity classification in context is much harder*  
   – if pursued at high degrees of fine-grainedness

2. Knowledge-based WSD may still be a way to go  
   – *but we need to employ true knowledge (bases)*

3. If distributional models perform better  
   – maybe *this knowledge is best captured in (structured) distributional models*  
     (see Ed Hovy’s talk)
Large-scale fine-grained NERC

FG-NERC
What is it? Why should we care?
How best to approach it?

Assessing the challenge of large-scale FG-NERC
First experiments on large-scale FG-NERC
Open issues and extensions

Distributional vs. knowledge-based accounts
Named Entities in Context

Conclusion
A case for corpus-based structured distributional models?
What to conclude?

Or maybe we should aim for acquiring the „perfect“ knowledge base, to enhance results?

Before doing such a big investment, we should test the chances of such an approach, by having the task solved by a perfect knowledge-based system: Let’s see how a human performs!
How do humans perform on this task?

Establish an upper bound for automatic FG-NE classification

Person sub-hierarchy from WordNet comprising 141 NE classes (synsets)

Annotation Task

Assign one (several) out of 141 NE classes to unknown person NEs in context (120 words) given the entire class hierarchy
How do humans perform on this task?
How do humans perform on this task?
How do humans perform on this task?
visitors will come to see the Gormleys in the next eighteen months – purchasing one million cups of tea at the very least and since each cup would generate 50 pence profit, I want the jobs and the money that these figures will create on a permanent basis. John, please can you help us? The Gormley statues are on sale for £1,500,000 and we simply must raise the money to keep this fantastic piece of art in this country. The wait3 function suspends execution of the current process until a child has exited, or until a signal is delivered whose action is to terminate the current process or to call a signal LABEL:

and Murcia were acquainted. Literes refers to the book as having been engraved (and presumably printed) in Antwerp. Murcia may have done the engraving himself. In the introductory letter to the reader he seems to be claiming credit personally for the attractive appearance of the tablature and says that the work was carried out abroad. (Note2).

LABEL:
Human Performance

Annotation set: Introductory contexts (w/o class label)
15 classes x up to 7 contexts
overall precision (exact match): 31.06%

<table>
<thead>
<tr>
<th>Class</th>
<th>Prec</th>
<th>Class</th>
<th>Prec</th>
<th>Class</th>
<th>Prec</th>
</tr>
</thead>
<tbody>
<tr>
<td>adventurer</td>
<td>33.33</td>
<td>journalist</td>
<td>0</td>
<td>priest</td>
<td>57.14</td>
</tr>
<tr>
<td>biologist</td>
<td>42.85</td>
<td>judge</td>
<td>14.28</td>
<td>ruler</td>
<td>14.28</td>
</tr>
<tr>
<td>diplomat</td>
<td>14.28</td>
<td>monk</td>
<td>33.33</td>
<td>scientist</td>
<td>71.42</td>
</tr>
<tr>
<td>engineer</td>
<td>42.85</td>
<td>physician</td>
<td>14.28</td>
<td>slave</td>
<td>16.66</td>
</tr>
<tr>
<td>entrepreneur</td>
<td>57.14</td>
<td>politician</td>
<td>14.28</td>
<td>surgeon</td>
<td>50.0</td>
</tr>
</tbody>
</table>

**Overall Precision:** 31.06 %
Human Performance

Annotation set: Follow-up contexts
15 classes x up to 14 contexts
overall precision (exact match): 28.01%

<table>
<thead>
<tr>
<th>Class</th>
<th>Prec</th>
<th>Class</th>
<th>Prec</th>
<th>Class</th>
<th>Prec</th>
</tr>
</thead>
<tbody>
<tr>
<td>adventurer</td>
<td>16.66</td>
<td>journalist</td>
<td>21.14</td>
<td>priest</td>
<td>14.28</td>
</tr>
<tr>
<td>biologist</td>
<td>21.14</td>
<td>judge</td>
<td>45.45</td>
<td>ruler</td>
<td>30.76</td>
</tr>
<tr>
<td>diplomat</td>
<td>25.0</td>
<td>monk</td>
<td>33.33</td>
<td>scientist</td>
<td>78.57</td>
</tr>
<tr>
<td>engineer</td>
<td>25.0</td>
<td>physician</td>
<td>42.85</td>
<td>slave</td>
<td>28.57</td>
</tr>
<tr>
<td>entrepreneur</td>
<td>0</td>
<td>politician</td>
<td>30.76</td>
<td>surgeon</td>
<td>0</td>
</tr>
</tbody>
</table>

Overall Precision: 28.01 %
Human Performance

Annotation set: All contexts
15 classes x up to 21 contexts
overall precision (exact match): 29.37%

<table>
<thead>
<tr>
<th>Class</th>
<th>Prec</th>
<th>Class</th>
<th>Prec</th>
<th>Class</th>
<th>Prec</th>
</tr>
</thead>
<tbody>
<tr>
<td>adventurer</td>
<td>22.22</td>
<td>journalist</td>
<td>14.28</td>
<td>priest</td>
<td>28.57</td>
</tr>
<tr>
<td>biologist</td>
<td>31.57</td>
<td>judge</td>
<td>35.29</td>
<td>ruler</td>
<td>25.0</td>
</tr>
<tr>
<td>diplomat</td>
<td>19.04</td>
<td>monk</td>
<td>33.33</td>
<td>scientist</td>
<td>76.19</td>
</tr>
<tr>
<td>engineer</td>
<td>31.57</td>
<td>physician</td>
<td>33.33</td>
<td>slave</td>
<td>23.07</td>
</tr>
<tr>
<td>entrepreneur</td>
<td>23.52</td>
<td>politician</td>
<td>25.0</td>
<td>surgeon</td>
<td>15.78</td>
</tr>
</tbody>
</table>

Overall Precision: 29.37 %
Human Performance

Annotation set: All contexts
15 classes x up to 21 contexts
overall precision (exact match): 29.37%

Surprising?
Only 10 points over the majority baseline (P 18.5%)
8 points below Extended Lesk (P 38.6%)
in contrast to MaxEnt-A&B (P 86.5%, introductory)

So it seems that even a „perfect“ knowledge base is not sufficient to solve the task successfully
What to conclude?

But then...

*Why do humans underperform compared to a MaxEnt classification system?*
What to conclude?

*Humans* should do better in *processing knowledge and making associations* than this is possible on basis of WordNet (using gloss overlap or Page Rank)

*Context snippets* are often *too unrestricted, or underspecified* to allow for an *informed decision*

*MACHINE LEARNERS*, if given enough and diverse examples, are better at *computing consistent association measures* to discriminate large sets of classes
Large-scale fine-grained NERC

FG-NERC
What is it? Why should we care?
How best to approach it?

Assessing the challenge of large-scale FG-NERC
First experiments on large-scale FG-NERC
Open issues and extensions

Distributional vs. knowledge-based accounts
Named Entities in Context

Conclusion
A case for corpus-based structured distributional models?
Conclusion

If this can be confirmed, we should indeed aim for a statistical semantics that exploits the law of great numbers

– and inject as much knowledge as we can in terms of structured distributional models

[...enhanced by contextual processing]
JUL:  
'Tis but thy name that is my enemy. (40)  
Thou art thyself, though not a Montague.  
What's Montague? it is nor hand, nor foot,  
Nor arm, nor face, nor any other part  
Belonging to a man. O, be some other name!  
What's in a name? That which we call a rose  
By any other name would smell as sweet.  
So Romeo would, were he not called Romeo  
Retain that dear perfection which he has  
Without that title. Romeo, do thou name me  
And for that name, which is no part of you,  
Take all myself.
Thank you