# An Incremental Entity-Mention Model for Coreference Resolution with Restrictive Antecedent Accessibility

Manfred Klenner, Don Tuggener

# Overview

- Coreference resolution for noun phrases (NEs, Nouns, Pers., Refl., Poss. Pronouns)
- Goal: Address drawbacks of the popular mention-pair model
- Empirical comparison of mention pair-model and incremental entity-mention model
- Real preprocessing (no gold standard information used, e.g treebanks / morphology)
- Simple salience measure based on grammatical functions vs. Machine Learning classifiers
- Evaluated on German and English corpora (SemEval'10, CoNLL, BioNLP)
- Focus on pair generation mechanics

# Starting point: Mention-pair model

Popular, often reimplemented architecture by e.g. Soon et al. (2001)[1]
How does it work? E.g. 3rd person pronouns

- Look for markables (potential antecedent NPs) within filter-defined window (distance restrictions, morphological agreement ...)
- Build pairs and their feature vectors:
  antecedent candidate $\leftrightarrow$ pronoun
- Binary classification: coreferent yes/no
- Positive pair with highest score according to ML classifier becomes antecedent
- Cluster positive pairs to establish coreference sets

[1] Soon, Wee Meng and Ng, Hwee Tou and Lim, Daniel Chung Yong (2001): A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Comput. Linguist., MIT Press, 27, 521-544*

# 3 Problems of the mention-pair model

**Redundancy**
All mentions of (implicitly) established coreference sets are accessible.

Positive binary classification of: $\textit{Hillary Clinton}_i \leftarrow \textit{she}_i$
Another $\textit{she}_?$ occurs. Results in two pairs:

$\textit{Hillary Clinton}_i \leftarrow \textit{she}_?$
$\textit{she}_i \leftarrow \textit{she}_?$

$\rightarrow$ same effect for positive and negative pairs

- Results in **huge training sets**, which are **imbalanced** (more negative examples due to pair generation mechanics)
- Leads to **classifiers biased towards negative classification**

# 3 Problems of the mention-pair model

**Transitivity**:
**Local perspective** of the mention-pair model, no means to enforce global constraints (directly).

*Angela Merkel*$_i$ ← *she*$_?$
*she*$_i$ ← *Hillary Clinton*$_?$

Classifier says yes to both pairs, clustering gives us the coreference set:
[Angela Merkel, she, Hillary Clinton]

But we know Hillary Clinton and Angela Merkel are exclusive.

→ **Additional optimisation** step is needed to **resolve such inconsistencies**, e.g. Klenner (2007)[2]

[2]Klenner, Manfred (2007): Enforcing Consistency on Coreference Sets. *Proceedings of the International Conference on Recent Advances in Natural Language Processing, 323-328*

# 3 Problems of the mention-pair model

**Underspecification** of antecedent candidates.

$Clinton_k \leftarrow she_?$

Is it good pair? Hillary or Bill Clinton?

# 3 Problems of the mention-pair model

A lot of (successful) research went into finding new/better features for the binary classifiers, c.f. Ng (2010)[3]

**Our focus** is **pair generation mechanics**: Progress in coreference resolution is possible by not only optimising classifier performance, but also the steps in the pipeline that determine what is presented to the classifier.

Deal with **Redundancy, Transitivity, Underspecification** in one step

$\longrightarrow$ Incremental entity-mention model

**Related work** on the entity-mention model has focussed on creating and exploring **cluster based features** (e.g. how many nouns are in the cluster etc.) $\longrightarrow$ improve classifiers

[3] Ng, Vincent (2010): Supervised Noun Phrase Coreference Research: The First Fifteen Years. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 1396-1411*

# Incremental Entity-Mention Model

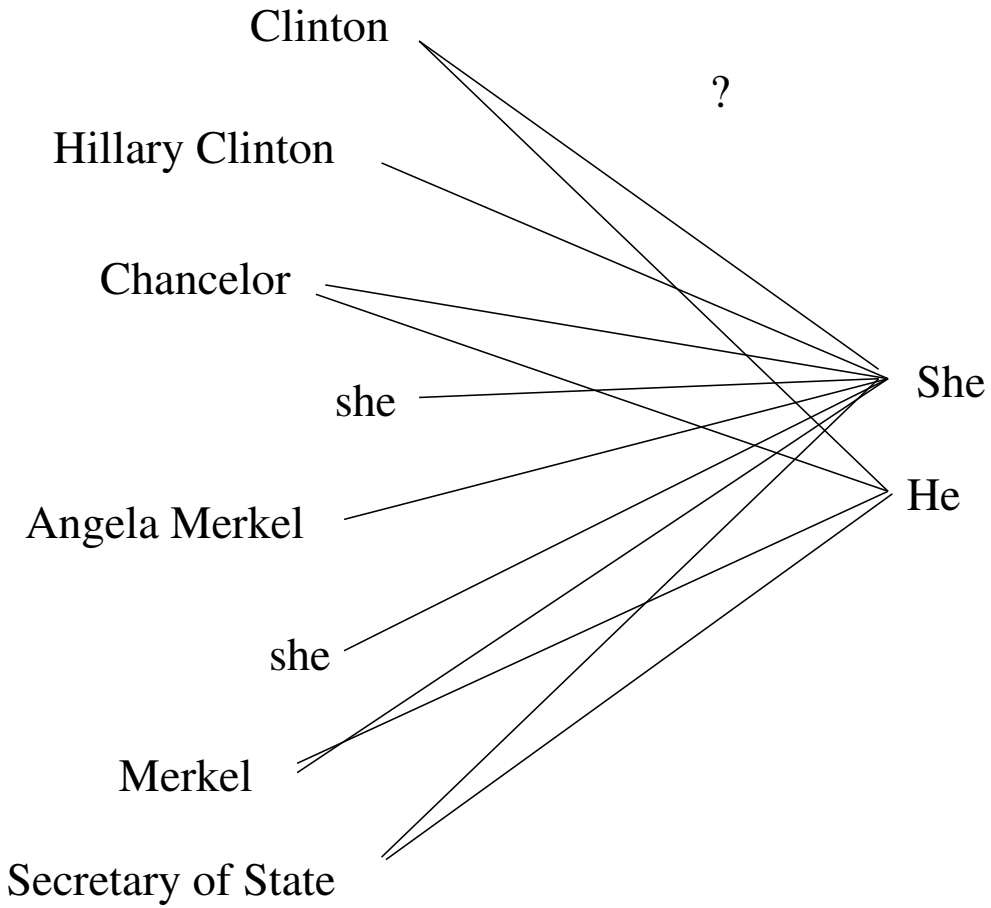
**Entity-mention model**: Evaluate candidate pair on the basis of emerging coreference sets
**Incremental**: Create/augment coreference sets "on the fly", right after classification

**Algorithm**:

- Process text from left to right, put first non-anaphoric NPs on waiting list (see e.g. Lappin & Leass (1994)[4]), as they might be valid antecedents
- Potentionally anaphoric NPs are compared to elements of the waiting list and a "virtual prototypes" per established coreference sets, which subsumes all the known morpho-syntactic and semantic information of the mentions of the set.
  e.g. *["Hillary Clinton", Fem, Sing, Person, ...]*
  - If we find a compatible antecedent, establish/append to coreference set
  - Else append to waiting list

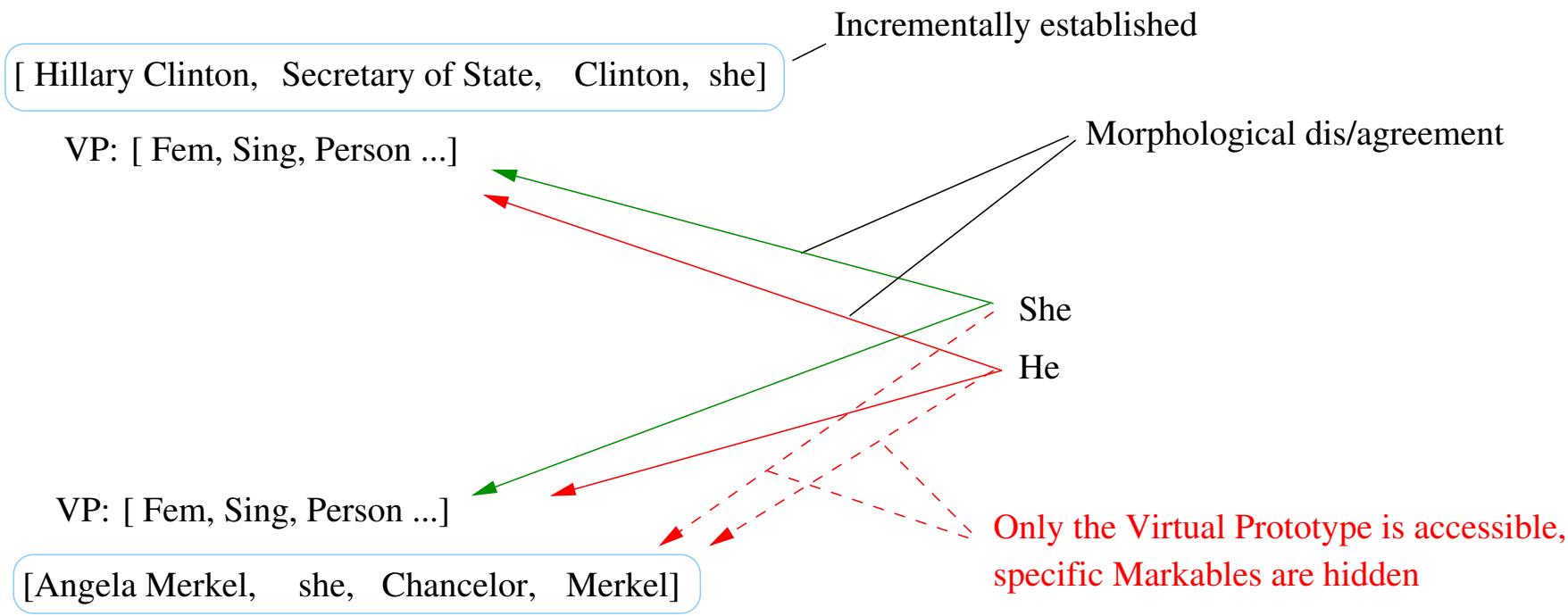
[4] Shalom Lappin and Herbert J Leass (1994): An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics, 20, 535-561*

# Antecedent Accessibility: Mention-pair model




–> 12 pairs for classification
–> Clustering needed
–> Resolve inconsistencies
[ Angela Merkel, Chancelor, He]

# Antecedent Accessibility: Entity-mention model




–> 2 pairs for classification:

[ Hillary Clinton, Fem, Sing, Person, ...] – She

[ Angela Merkel, Fem, Sing, Person , ...] – She

# How do we solve the three problems? Redundancy, Transitivity, Underspecification

**Redundancy**

Instead of:
*“Hillary Clinton”*$_i$ $\leftarrow$ *she*$_?$
*she*$_i$ $\leftarrow$ *she*$_?$

we do:
(“Hillary Clinton”, she)[Fem, Sing, Person]$_i$ $\leftarrow$ *she*$_?$

$\rightarrow$ This **reduces the number of considered pairs from the cardinality of a set to 1**, for both positive and negative examples

# How do we solve the three problems? Redundancy, Transitivity, Underspecification

**Transitivity**

E.g. **Binding theory**

- *“Clinton had met her before.”*.
  *Clinton* and *her* cannot be coreferent (**c-command**)
- all mentions of the so far established *Clinton* coreference set, say [“Hillary Clinton”, she, Clinton] are **transitively exclusive** and need not be considered as antecedent candidates for *“her”*.

Or cataphora

- i.e. [“Angela Merkel”, she] $\rightarrow$ *“Hillary Clinton”* cannot be coreferent with this “she”, because *“Angela Merkel”* and *“Hillary Clinton”* are exclusive.

# How do we solve the three problems? Redundancy, Transitivity, Underspecification

**Underspecification**:

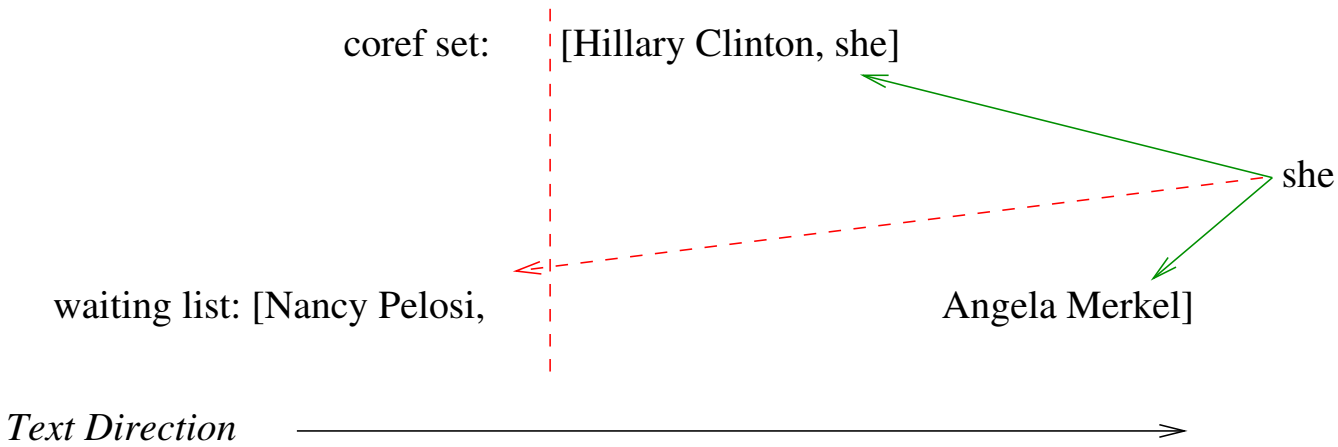
$Clinton_k \leftarrow she_?$

Is it good pair?

Yes, if $Clinton_k$ is part of the set [“Hillary Clinton”, Clinton]
No, if $Clinton_k$ is part of the set [“Bill Clinton”, Clinton]

# Conclusions

- Not **storing** and making available mentions but **discourse units**
- Moving towards a model that, at best, takes into account **human cognitive capacity** and its limitations
- Link to cognitive science? People do not store individual mentions for reference but discourse units in memory
- Furthermore, access to markables on waiting list can be restricted. Only accessible if, e.g. more recent than compatible coreference set candidates



# Evaluation: Reducing antecedent candidates

Comparing mention-pair and entity-mention training instance numbers
Training instances TüebaD/Z (Fold 1 of 5)
**Reduction by a factor of 4**

| Anaphora Type | Pos | Neg |
|---|---|---|
| *Mention-pair model* (**171526 instances**) | | |
| Nouns | 5626 | 5144 |
| Relative pronouns | 1428 | 2459 |
| Reflexive pronouns | 1372 | 728 |
| Possessive pronouns | **5346** | **21571** |
| Personal pronouns | **23025** | **104827** |
| Total | 36797 | 134729 |
| *Entity-mention model* (**40229 instances**) | | |
| Nouns | 1776 | 3787 |
| Relative pronouns | 1382 | 2330 |
| Reflexive pronouns | 462 | 530 |
| Possessive pronouns | **1416** | **8156** |
| Personal pronouns | **4023** | **16367** |
| Total | 9059 | 31170 |

# Evaluation: Model comparison

- TübaD/Z corpus (German), 5 Fold cross-evaluation
- TiMBL (based on k-NN)
- Standard feature set, e.g. Soon (2001)
- Classifier and feature set per markable type (manual tuning)
- Filtering based on markable type
- Same feature and filter set for both models
- Own implementation of $\text{CEAF}_m$ (Luo, 2005)[5] $\rightarrow$ we disregard singletons

[5] Luo, Xiaoqiang (2005): On Coreference Resolution Performance Metrics. *HLT '05: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 25-32*

# Evaluation: Model comparison

| Model | F1 | P | R |
|---|---|---|---|
| Mention-pair (TiMBL + ILP) | 49.35 | **53.67** | 45.69 |
| Entity-mention (TiMBL) | **52.79** | 52.88 | **52.70** |

- Huge gain in Recall (+ 7.01%)
- Lose a bit of Precision (- 0.79%)
- F1 significantly increased (+ 3.44%)
- Huge gap in processing time
- Results are low $\rightarrow$ coreference resolution with real preprocessing is hard!

# Error Analysis

| | Without filtering | With filtering |
|---|---|---|
| | F1 | F1 |
| Nouns | **72.70** | **62.61** |
| Pers. pron. | 60.42 | 58.86 |
| Rel. pron. | 56.25 | 55.97 |
| Poss. pron. | 56.06 | 55.81 |
| Refl. pron. | 55.68 | 54.16 |
| System | - | **53.86** |

- We run the system on fold 1 and **resolve one markable type perfectly** (gold information)
- E.g. how good would the system be, if we resolved Nouns (1st row) perfectly?
- How good are the filters? Do they delete true mentions?
- 19% loss due to nominal markables

# Shared Task Evaluations: A simple salience measure

- Based on grammatical functions
- Salience of "subject": $\frac{\textit{Number of true mentions bearing subject}}{\textit{Total number of true mentions}}$
- Hierarchical ordering of grammatical functions (subject $>$ object $>$ ...)
- No training, really fast, easy to adapt (no feature vectors)

| Model | F1 | P | R |
|---|---|---|---|
| Mention-pair (TiMBL + ILP) | 49.35 | **53.67** | 45.69 |
| Entity-mention (TiMBL) | **52.79** | 52.88 | **52.70** |
| Entity-mention (salience) | 51.41 | 52.03 | 50.82 |

# SemEval ’10: Own post task evaluation

- SemEval ’10: Coreference resolution in multiple languages. Comparing evaluation metrics.
- “open regular”: Real preprocessing, additional ressources and tools can be used
- Except for MUC (English) best results
- Overall strong Precision compared to other systems

| | CEAF | | | MUC | | | BCUB | | | BLANC | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| System | R | P | F1 | R | P | F1 | R | P | F1 | R | P | F1 |
| German, open regular | | | | | | | | | | | | |
| bart | 61.4 | 61.2 | 61.3 | **61.4** | 36.1 | 45.5 | 75.3 | 58.3 | 65.7 | **55.9** | 60.3 | 57.3 |
| **incr** | **76.8** | **70.4** | **73.4** | 50.4 | **47.1** | **48.7** | **81.7** | **75.6** | **78.5** | 55 | **72.6** | **57.8** |
| English, open regular | | | | | | | | | | | | |
| bart | 70.1 | 64.3 | 67.1 | **62.8** | 52.4 | 57.1 | 74.9 | 67.7 | 71.1 | 55.3 | 73.2 | 57.7 |
| corry-b | 70.4 | 67.4 | 68.9 | 55.0 | 54.2 | 54.6 | 73.7 | 74.1 | 73.9 | 57.1 | 75.7 | 60.6 |
| corry-c | **70.9** | 67.9 | 69.4 | 54.7 | 55.5 | 55.1 | 73.8 | 73.1 | 73.5 | 57.4 | 63.8 | 59.4 |
| corry-m | 66.3 | 63.5 | 64.8 | 61.5 | 53.4 | **57.2** | **76.8** | 66.5 | 71.3 | **58.5** | 56.2 | 57.1 |
| **incr** | 67.6 | **73** | **70.2** | 34 | **62.5** | 44.1 | 66.7 | **86** | **75.1** | 57.1 | **78.4** | **61.1** |

# BioNLP 2011

| Team | R | P | F1 |
|---|---|---|---|
| A | 22.18 | 73.26 | 34.05 |
| **incr** | 21.48 | 55.45 | 30.96 |
| B | 19.37 | 63.22 | 29.65 |
| C | 14.44 | 67.21 | 23.77 |
| D | 3.17 | 3.47 | 3.31 |
| E | 0.70 | 0.25 | 0.37 |

- Task: Finding anaphoric references to proteins and genes.
- Collaboration with Ontogene group[6]
- Only antecedent-anaphora pairs are considered in evaluation and not full coreference sets. Anaphora resolution still benefits from establishing them to restrict antecedent accessibility (enforce transitivity constraints).

[6] http://www.ontogene.org/

# CoNLL 2011 shared task

| Metric | R | P | F1 |
|---|---|---|---|
| CEAFM | 51.08 | 51.08 | 51.08 |
| CEAFE | 44.35 | 39.93 | 42.03 |
| BCUB | 60.91 | 70.69 | 65.44 |
| BLANC | 63.63 | 72.58 | 66.81 |
| MUC | 45.18 | 49.83 | 47.39 |

- CoNLL ranking: 51.77, 4th of 5 in the open track (we use Wikipedia for NE classification)
- Closed track: 18 participants, average: 50.09 (best 57.79, last 31.28)

$\rightarrow$ Average results

# Conclusion

- Easy to implement architecture: no clustering, no enforcing of transitive consistency
- Restrict antecedent candidate accessibility: Discourse units only, Binding theory, linguistic filtering, waiting list access (more recent or more salient than coreference set candidates only)
- Claim: closer to human coreference/text processing than mention-pair model
- Fast, even faster with simple salience measure based on grammatical functions (no training)
- Coreference resolution with real preprocessing still very hard (see CoNLL results in general)! Challenge: Nominal anaphora (non string matching)
- Web demos available: http://kitt.cl.uzh.ch/kitt/coref/