Balancing multiple models for coreference resolution using integer programming

Pascal Denis

EPI Alpage
INRIA/Université Paris 7
pascal.denis@inria.fr
(Joint work with Jason Baldridge, UT Austin)

Computational Linguistics colloquium
Universität Heidelberg
July 2, 2009
Coreference Resolution (CR)

Task definition
To partition the set of mentions into equivalence classes (or chains) that correspond to discourse entities

Example
Clinton told National Public Radio that his answers to questions about Lewinsky were constrained by Starr’s investigation. NPR reporter Mara Liasson asked Clinton "whether you had any conversations with her about her testimony, had any conversations at all."
Coreference Resolution (CR)

Task definition
To partition the set of mentions into equivalence classes (or chains) that correspond to discourse entities

Example
Clinton told National Public Radio that his answers to questions about Lewinsky were constrained by Starr’s investigation. NPR reporter Mara Liasson asked Clinton "whether you had any conversations with her about her testimony, had any conversations at all."
Difficult and critical task

Challenges
- Numerous information sources at play: linguistic and non-linguistic (AI-complete problem!)
- Most (if not all) of these sources are not always reliable
- Different referring expressions show different resolution strategies (cf. Ariel 1988, Gundel et al. 1994)

Applications
- Information Extraction
- Text Summarization
- Question Answering
Trends in Anaphora/Coreference Resolution

- **Knowledge-based systems**: Hobbs 1978, BFP 1987
  - Brittle, hard to build, poor evaluation
  - Knowledge-lean and more robust, but “fiddly”
- **Machine learning systems**: Soon et al, 2001; Ng & Cardie (2002a); Kehler et al. 2004
  - More robust, easy to build, better performance
Standard machine learning approach

Standard approach works in two steps:

1. **Classification**: decide whether *pairs* of mentions corefer or not
2. **Clustering**: merge corefential pairs into *chains* (typically, closest-first/best-first clustering)

Still a lot of room for improvement: 60-70% in $f$-score
Current Limitations

- **Knowledge-bottleneck**
  - Numerous information sources at play, but systems often rely on small sets of shallow features
  - Predicting and incorporating richer information is challenging due to error propagation (e.g., Ng & Cardie, 2002a, Kehler et al. 2004)

- **Decision Locality**
  - Clustering decisions are *locally* optimized, but they depend on one another:

```
Mr. Clinton  Clinton  she
   .7        .62      .12
```
Mentions can be divided into two classes:

- Mentions that are linked to an existing entity (="discourse-old" or "anaphora")
- Mentions that “introduce” a new entity (="discourse-new" or "non-anaphors")

Many errors originate from anaphoricity errors:

- True anaphora fail to be resolved
- True non-anaphors are resolved

Ideally, one would like to resolve all and only the true anaphora
Additional knowledge sources (II)

Named entity type

- Mentions are instances of different semantic types: e.g., PERSON, ORGANIZATION, LOCATION, GPE, FACILITY
- Some coreference errors can be traced back to semantic types mismatches
- Ideally, we want to make sure that two mentions can be coreferential only if they have the same entity type
Pipeline Model
Ng & Cardie (2002b), Ng (2004)

- Learn **separate classifiers** for coreference, discourse status, named entity classification
  - \( cl_{\text{coref}} : \mathcal{M} \times \mathcal{M} \rightarrow \{\text{COREF, NOT-COREF}\} \)
  - \( cl_{ds} : \mathcal{M} \rightarrow \{\text{OLD, NEW}\} \)
  - \( cl_{ne} : \mathcal{M} \rightarrow \{\text{PER, ORG, LOC, GPE, FAC}\} \)

- The discourse status and named entity classification models are used as **filters** during test:
  - only mentions classified \( \text{OLD} \) are considered by the coreference model
  - for a candidate anaphor of type \( t \), only mentions of type \( t \) are considered as candidate antecedent
The “curse” of the pipeline

Example

Clinton told National Public Radio that his answers to questions about Lewinsky were constrained by Starr’s investigation. NPR reporter Mara Liasson asked Clinton "whether you had any conversations with her about her testimony, had any conversations at all."

NE Model

\[
\langle \text{“Clinton”}, \text{PER} \rangle \quad .83 \\
\langle \text{“Nat’l Public Radio”}, \text{ORG} \rangle \quad .61 \\
\langle \text{“NPR”}, \text{GPE} \rangle \quad .59 \\
\ldots \\
\ldots
\]

Coreference Model

\[
\langle \text{“Clinton”}, \text{“Nat’l Public Radio”} \rangle \quad .02 \\
\langle \text{“Clinton”}, \text{“his”} \rangle \quad .82 \\
\langle \text{“Clinton”}, \text{“Lewinsky”} \rangle \quad .01 \\
\langle \text{“Nat’l Public Radio”}, \text{“NPR”} \rangle \quad .92 \\
\ldots \\
\ldots
\]
Ng and Cardie (2002b) report overall decrease (due to important drops in recall) when using a discourse status classifier.

Improvements with the pipeline require careful tuning of “anaphoricity” threshold on held-out data (Ng, 2004).

Coreference resolution and discourse status determination are codependent tasks: they should be modelled as a joint problem.
To avoid globally incoherent chains, we would like to be able to enforce *transitivity* on the final decisions of the models.
To avoid globally incoherent chains, we would like to be able to enforce transitivity on the final decisions of the models.
Inference with classifiers

- Train local models for different subtasks
- But impose constraints on the classifiers’ labels ensuring that final decisions are mutually consistent
- Constraints can relate different classifiers labels, but also labels from the same classifier
- Goal: find the best global assignment, that maximizes the local probabilities while satisfying the constraints
- Inference can be modelled as an optimization problem, cast in an integer linear programming (ILP) formulation
Why ILP?

- Expressive: can represent many types of constraints in declarative fashion
- Optimal: guarantee to find the optimal solution
- Fast: existing packages (CPLEX, lpsolve, GLPK) are able to quickly solve very large problems
- ILP works well for number of NLP tasks:
  - entity and relation identification (Roth & Yih, 2004)
  - sentence aggregation for generation (Barzilay & Lapata, 2006)
- Previous work on coreference: Denis & Baldridge 2007, Klenner 2008

P. Denis
Balancing multiple models for coref. resolution using ILP
Rest of the talk

1. Base Models
   - Coreference classifier
   - Discourse status classifier
   - Named entity classifier
   - Data and Evaluation
   - Pipeline experiments

2. ILP formulations
   - Simple formulation
   - Joint formulations
   - Adding global constraints

3. Summary and future work
1. **Base Models**
   - Coreference classifier
   - Discourse status classifier
   - Named entity classifier
   - Data and Evaluation
   - Pipeline experiments

2. **ILP formulations**
   - Simple formulation
   - Joint formulations
   - Adding global constraints

3. **Summary and future work**
Rest of the talk

1. Base Models
   - Coreference classifier
   - Discourse status classifier
   - Named entity classifier
   - Data and Evaluation
   - Pipeline experiments

2. ILP formulations
   - Simple formulation
   - Joint formulations
   - Adding global constraints

3. Summary and future work
Model #1: Coreference classifier

Modelling

We estimate the probability, \( P_C(\text{COREF} | \langle i, j \rangle) \), of a pair of mentions \( \langle i, j \rangle \) being coreferential.

Maxent formulation

\[
P_C(\text{COREF} | \langle i, j \rangle) = \frac{\exp\left( \sum_{k=1}^{n} \lambda_k f_k(\langle i, j \rangle, \text{COREF}) \right)}{\sum_{c \in \{\text{COREF}, \neg\text{COREF}\}} \exp\left( \sum_{k=1}^{n} \lambda_k f_k(\langle i, j \rangle, c) \right)}
\]
The text is scanned L-to-R and for each anaphoric mention $j$:

- A **positive instance** is created between $j$ and its *closest* antecedent $i$
- **Negative instances** are created between $j$ and all the mentions $k$ intervening between $i$ and $j$
Testing procedure
Soon et al. (2001)

The text is scanned L-to-R and for each mention $i$:

- A test instance is created between $i$ and every mention $j$ that precedes it
- Each instance is presented to the classifier, which returns a probability
- The process terminates as soon as an antecedent is found (a probability $> .5$) or the beginning of the text is reached

NB: only one antecedent per anaphor!
Feature set and parameter estimation

**Features**
- Anaphor: NP type, NE, length, POS context, etc.
- Antecedent candidate: NP type, NE, POS context, etc.
- Relational: distance, (sub-)string matching, morphosyntactic agreement, alias, appositive, etc.

**Parameter estimation**
- Limited memory variable metric algorithm
- Prior with variance of 1000
Rest of the talk

1. Base Models
   - Coreference classifier
   - Discourse status classifier
   - Named entity classifier
   - Data and Evaluation
   - Pipeline experiments

2. ILP formulations
   - Simple formulation
   - Joint formulations
   - Adding global constraints

3. Summary and future work

P. Denis

Balancing multiple models for coref. resolution using ILP
Model #2: Discourse status classifier

Modelling

We estimate the probability, $P_{DS}(OLD|i)$, of having a mention $i$ being anaphoric.

Maxent formulation

$$P_{DS}(OLD|i) = \frac{\exp\left(\sum_{k=1}^{n} \lambda_k f_k(i, OLD)\right)}{\sum_{c \in \{OLD, NEW\}} \exp\left(\sum_{k=1}^{n} \lambda_k f_k(i, c)\right)}$$
Feature set and parameter estimation

Features
- Mention-based: NP type, length, etc.
- Contextual: position in sentence, count of previous identical mentions, etc.

Parameter estimation
- Limited memory variable metric algorithm
- Prior with variance of 1000

Our anaphoricity classifier has accuracy of 80.9%
Rest of the talk

1. Base Models
   - Coreference classifier
   - Discourse status classifier
   - Named entity classifier
   - Data and Evaluation
   - Pipeline experiments

2. ILP formulations
   - Simple formulation
   - Joint formulations
   - Adding global constraints

3. Summary and future work
Model #3: Named entity classifier

**Modelling**
We estimate the probability, $P_{DS}(t|i)$, of having a mention $i$ being of type $t$, where $t \in \{\text{PER, ORG, LOC, GPE, FAC}\}$

**Maxent formulation**

$$P_{ne}(t|i) = \frac{\exp \sum_{j=1}^{m} \lambda_{j} f_{j}(i, t)}{\sum_{t'\in\{\text{PER, ORG, LOC, GPE, FAC}\}} \exp \sum_{j=1}^{m} w_{j} f_{j}(i, t')}$$
Feature set and parameter estimation

Features
- Mention string: capitalization, punctuation, head word, etc.
- Wordnet senses: incl. senses in the hypernym closure
- Context: words and POSs

Parameter estimation
- Limited memory variable metric algorithm
- Prior with variance of 1000

Our NE classifier has accuracy of 79.5%
1. **Base Models**
   - Coreference classifier
   - Discourse status classifier
   - Named entity classifier
   - Data and Evaluation
   - Pipeline experiments

2. **ILP formulations**
   - Simple formulation
   - Joint formulations
   - Adding global constraints

3. **Summary and future work**
ACE Phase 2 Corpus

- Corpus annotated with coreference chains
- The mentions are restricted to 5 types of entities: FACility, GPE (geo-political entity), LOCATION, ORGANIZATION, PERSON,
- Data is split into three datasets: BNEWS, NPAPER, and NWIRE
- Each dataset is divided into a train part (~ 10000 mentions) and a test part (~ 2500 mentions)
How well do the predicted entities $S$ match the true entities $T$?

- **MUC metric**: compute the overall number of **links** that are common to $S$ and $T$.
- **B³ metric**: compute for each **mention** the number of mentions that are correctly linked to it.
- **CEAF metric**: compute the best one-to-one mapping between the **chains** $S$ and $T$. 

P. Denis

Balancing multiple models for coref. resolution using ILP
The most widely used, the MUC metric has important flaws (Bagga & Baldwin 1998; Luo 2005):

- Singleton chains are ignored (no links)
- Gives high recall for large chains without precision penalty.

Consider a ridiculous baseline system: every document has a single entity. Results on ACE Phase 2:

<table>
<thead>
<tr>
<th></th>
<th>R</th>
<th>P</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUC</td>
<td>100.0</td>
<td>60.2</td>
<td>75.1</td>
</tr>
<tr>
<td>CEAF</td>
<td>20.2</td>
<td>20.2</td>
<td>20.2</td>
</tr>
</tbody>
</table>
Rest of the talk

1. **Base Models**
   - Coreference classifier
   - Discourse status classifier
   - Named entity classifier
   - Data and Evaluation
   - Pipeline experiments

2. **ILP formulations**
   - Simple formulation
   - Joint formulations
   - Adding global constraints

3. **Summary and future work**
Baseline experiments

4 systems

1. COREF$_{base}$: base pairwise coreference classifier
2. CASCADE$_{a\rightarrow c}$: only mentions classified as anaphoric by the DS model are resolved
3. CASCADE$_{e\rightarrow c}$: only mentions given the same type label as the anaphor are part of possible antecedents
4. CASCADE$_{a,e\rightarrow c}$: CASCADE$_{a\rightarrow c}$ and CASCADE$_{e\rightarrow c}$
## Experiment #1: Pipeline results

<table>
<thead>
<tr>
<th>System</th>
<th>MUC</th>
<th></th>
<th>B³</th>
<th></th>
<th>CEAF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F</td>
<td>R</td>
<td>P</td>
</tr>
<tr>
<td>COREF&lt;sub&gt;base&lt;/sub&gt;</td>
<td>60.8</td>
<td>72.6</td>
<td>66.2</td>
<td>62.4</td>
<td>77.7</td>
</tr>
<tr>
<td>CASCADE&lt;sub&gt;a→c&lt;/sub&gt;</td>
<td>64.9</td>
<td>72.3</td>
<td>68.4</td>
<td>65.6</td>
<td>74.1</td>
</tr>
<tr>
<td>CASCADE&lt;sub&gt;e→c&lt;/sub&gt;</td>
<td>56.3</td>
<td>75.2</td>
<td>64.4</td>
<td>59.6</td>
<td>82.4</td>
</tr>
<tr>
<td>CASCADE&lt;sub&gt;a,e→c&lt;/sub&gt;</td>
<td>61.3</td>
<td>68.8</td>
<td>64.8</td>
<td>62.5</td>
<td>73.8</td>
</tr>
<tr>
<td>ORACLE&lt;sub&gt;a,e&lt;/sub&gt;</td>
<td>83.2</td>
<td>83.2</td>
<td>83.2</td>
<td>79.0</td>
<td>78.2</td>
</tr>
</tbody>
</table>
## Experiment #1: Pipeline results

<table>
<thead>
<tr>
<th>System</th>
<th>MUC</th>
<th>B³</th>
<th>CEAF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F</td>
</tr>
<tr>
<td>COREF&lt;sub&gt;base&lt;/sub&gt;</td>
<td>60.8</td>
<td>72.6</td>
<td>66.2</td>
</tr>
<tr>
<td>CASCADE&lt;sup&gt;a→c&lt;/sup&gt;</td>
<td>64.9</td>
<td>72.3</td>
<td>68.4</td>
</tr>
<tr>
<td>CASCADE&lt;sup&gt;e→c&lt;/sup&gt;</td>
<td>56.3</td>
<td>75.2</td>
<td>64.4</td>
</tr>
<tr>
<td>CASCADE&lt;sup&gt;a,e→c&lt;/sup&gt;</td>
<td>61.3</td>
<td>68.8</td>
<td>64.8</td>
</tr>
<tr>
<td>ORACLE&lt;sup&gt;a,e&lt;/sup&gt;</td>
<td>83.2</td>
<td>83.2</td>
<td>83.2</td>
</tr>
</tbody>
</table>

P. Denis

Balancing multiple models for coref. resolution using ILP
### Experiment #1: Pipeline results

<table>
<thead>
<tr>
<th>System</th>
<th>MUC</th>
<th>B³</th>
<th>CEAF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F</td>
</tr>
<tr>
<td>COREF&lt;sub&gt;base&lt;/sub&gt;</td>
<td>60.8</td>
<td>72.6</td>
<td>66.2</td>
</tr>
<tr>
<td>CASCADE&lt;sub&gt;a→c&lt;/sub&gt;</td>
<td>64.9</td>
<td>72.3</td>
<td>68.4</td>
</tr>
<tr>
<td>CASCADE&lt;sub&gt;e→c&lt;/sub&gt;</td>
<td>56.3</td>
<td>75.2</td>
<td>64.4</td>
</tr>
<tr>
<td>CASCADE&lt;sub&gt;a,e→c&lt;/sub&gt;</td>
<td>61.3</td>
<td>68.8</td>
<td>64.8</td>
</tr>
<tr>
<td>ORACLE&lt;sub&gt;a,e&lt;/sub&gt;</td>
<td>83.2</td>
<td>83.2</td>
<td>83.2</td>
</tr>
</tbody>
</table>
## Experiment #1: Pipeline results

<table>
<thead>
<tr>
<th>System</th>
<th>MUC</th>
<th>B³</th>
<th>CEAF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F</td>
</tr>
<tr>
<td>COREF&lt;sub&gt;base&lt;/sub&gt;</td>
<td>60.8</td>
<td>72.6</td>
<td><strong>66.2</strong></td>
</tr>
<tr>
<td>CASCADE&lt;sub&gt;a→c&lt;/sub&gt;</td>
<td>64.9</td>
<td>72.3</td>
<td><strong>68.4</strong></td>
</tr>
<tr>
<td>CASCADE&lt;sub&gt;e→c&lt;/sub&gt;</td>
<td>56.3</td>
<td><strong>75.2</strong></td>
<td>64.4</td>
</tr>
<tr>
<td>CASCADE&lt;sub&gt;a,e→c&lt;/sub&gt;</td>
<td>61.3</td>
<td>68.8</td>
<td><strong>64.8</strong></td>
</tr>
<tr>
<td>ORACLE&lt;sub&gt;a,e&lt;/sub&gt;</td>
<td>83.2</td>
<td>83.2</td>
<td><strong>83.2</strong></td>
</tr>
<tr>
<td>System</td>
<td>MUC</td>
<td>B³</td>
<td>CEAF</td>
</tr>
<tr>
<td>--------------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F</td>
</tr>
<tr>
<td>COREF&lt;sub&gt;base&lt;/sub&gt;</td>
<td>60.8</td>
<td>72.6</td>
<td>66.2</td>
</tr>
<tr>
<td>CASCADE&lt;sub&gt;a→c&lt;/sub&gt;</td>
<td>64.9</td>
<td>72.3</td>
<td>68.4</td>
</tr>
<tr>
<td>CASCADE&lt;sub&gt;e→c&lt;/sub&gt;</td>
<td>56.3</td>
<td>75.2</td>
<td>64.4</td>
</tr>
<tr>
<td>CASCADE&lt;sub&gt;a,e→c&lt;/sub&gt;</td>
<td>61.3</td>
<td>68.8</td>
<td>64.8</td>
</tr>
<tr>
<td>ORACLE&lt;sub&gt;a,e&lt;/sub&gt;</td>
<td>83.2</td>
<td>83.2</td>
<td>83.2</td>
</tr>
</tbody>
</table>
Rest of the talk

1. **Base Models**
   - Coreference classifier
   - Discourse status classifier
   - Named entity classifier
   - Data and Evaluation
   - Pipeline experiments

2. **ILP formulations**
   - Simple formulation
   - Joint formulations
   - Adding global constraints

3. **Summary and future work**
ILP basics

ILP problem

Optimization problem where:

- the **objective function** can be specified as a *linear* function of certain variables:
  \[ c_1 x_1 + c_2 x_2 + \ldots + c_n x_n \] (\(c_i\) are assignment costs)

- the **constraints** can be formulated as *equalities* or *inequalities* on those variables:
  \[ a_1 x_1 + a_2 x_2 + \ldots + a_n x_n \leq b \]

- values for the variables are **integers**
Rest of the talk

1. Base Models
   - Coreference classifier
   - Discourse status classifier
   - Named entity classifier
   - Data and Evaluation
   - Pipeline experiments

2. ILP formulations
   - Simple formulation
   - Joint formulations
   - Adding global constraints

3. Summary and future work

P. Denis
Balancing multiple models for coref. resolution using ILP
CR as an ILP problem: first try

- **M**: Set of mentions in document
- **P**: Set of mention pairs: \( \langle i, j \rangle \in M \times M, i < j \)
- **\( x_{\langle i, j \rangle} \)**: Coreference variables: 
  \[
  x_{\langle i, j \rangle} = \begin{cases} 
  1 & \text{if } i \text{ and } j \text{ corefer} \\
  0 & \text{otherwise} 
  \end{cases}
  \]
- **\( c^C \)**: Coreference cost function: 
  \[
  c^C_{\langle i, j \rangle} = P_{C_l}(\text{COREF}|\langle i, j \rangle)
  \]

**ILP problem**

- maximize: 
  \[
  \sum_{\langle i, j \rangle \in P} c^C_{\langle i, j \rangle} \cdot x_{\langle i, j \rangle} + (1 - c^C_{\langle i, j \rangle}) \cdot (1 - x_{\langle i, j \rangle})
  \]

- subject to: 
  \[x_{\langle i, j \rangle} \in \{0, 1\} \quad \forall \langle i, j \rangle \in P\]
Experiment #2: \( \text{COREF}_{\text{base}} \) vs. Simple ILP formulation

<table>
<thead>
<tr>
<th>System</th>
<th>MUC</th>
<th>B(^3)</th>
<th>CEAF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F</td>
</tr>
<tr>
<td>( \text{COREF}_{\text{base}} )</td>
<td>60.8</td>
<td>72.6</td>
<td>66.2</td>
</tr>
<tr>
<td>( \text{ILP}_c )</td>
<td>70.3</td>
<td>72.7</td>
<td>71.5</td>
</tr>
</tbody>
</table>

\( \text{ILP}_c \) is equivalent to taking all the links with probability \( > .5 \) (potentially more than one antecedent!)
## Experiment #2: COREF<sub>base</sub> vs. Simple ILP formulation

<table>
<thead>
<tr>
<th>System</th>
<th>MUC</th>
<th>B&lt;sup&gt;3&lt;/sup&gt;</th>
<th>CEAF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F</td>
</tr>
<tr>
<td>COREF&lt;sub&gt;base&lt;/sub&gt;</td>
<td>60.8</td>
<td>72.6</td>
<td>66.2</td>
</tr>
<tr>
<td>ILP&lt;sub&gt;c&lt;/sub&gt;</td>
<td>70.3</td>
<td>72.7</td>
<td>71.5</td>
</tr>
</tbody>
</table>

ILP<sub>c</sub> is equivalent to taking all the links with probability > .5 (potentially more than one antecedent!)
1. Base Models
   - Coreference classifier
   - Discourse status classifier
   - Named entity classifier
   - Data and Evaluation
   - Pipeline experiments

2. ILP formulations
   - Simple formulation
   - Joint formulations
   - Adding global constraints

3. Summary and future work
Joint Formulations (I)

Integrating DS model

- $\mathcal{M}$: Set of mentions in document
- $P$: Set of mention pairs: $\langle i, j \rangle \in \mathcal{M} \times \mathcal{M}, i < j$
- $x_{\langle i, j \rangle}$: Coreference variables: $x_{\langle i, j \rangle} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ corefer} \\ 0 & \text{otherwise} \end{cases}$
- $c^C$: Coreference Cost function: $c_{\langle i, j \rangle} = P_{Cl}(\text{COREF}|\langle i, j \rangle)$
- $y_j$: Anaphoricity variables: $y_j = \begin{cases} 1 & \text{if } j \text{ is anaphoric} \\ 0 & \text{otherwise} \end{cases}$
- $c^A$: Anaphoricity Cost function: $c^A_j = P_{A}(\text{ANAPH}|j)$
Joint Formulations (I)

Integrating DS model

\( \mathcal{M} \) Set of mentions in document

\( P \) Set of mention pairs: \( \langle i, j \rangle \in \mathcal{M} \times \mathcal{M}, i < j \)

\( x_{\langle i,j \rangle} \) Coreference variables: \( x_{\langle i,j \rangle} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ corefer} \\ 0 & \text{otherwise} \end{cases} \)

\( c^C \) Coreference Cost function: \( c_{\langle i,j \rangle} = P_{Cl}(\text{COREF}|\langle i, j \rangle) \)

\( y_j \) Anaphoricity variables: \( y_j = \begin{cases} 1 & \text{if } j \text{ is anaphoric} \\ 0 & \text{otherwise} \end{cases} \)

\( c^A \) Anaphoricity Cost function: \( c^A_j = P_A(\text{ANAPH}|j) \)
Joint Formulations (I)

**Integrating DS model**

- **\( \mathcal{M} \)**: Set of mentions in document
- **\( P \)**: Set of mention pairs: \( \langle i, j \rangle \in \mathcal{M} \times \mathcal{M}, i < j \)
- **\( x_{\langle i, j \rangle} \)**: Coreference variables: \( x_{\langle i, j \rangle} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ corefer} \\ 0 & \text{otherwise} \end{cases} \)
- **\( c^C \)**: Coreference Cost function: \( c_{\langle i, j \rangle} = P_{CI}(\text{COREF}|\langle i, j \rangle) \)
- **\( y_j \)**: Anaphoricity variables: \( y_j = \begin{cases} 1 & \text{if } j \text{ is anaphoric} \\ 0 & \text{otherwise} \end{cases} \)
- **\( c^A \)**: Anaphoricity Cost function: \( c^A_j = P_A(\text{ANAPH}|j) \)
Integrating DS model

maximize: \[ \sum_{\langle i, j \rangle \in P} c_{\langle i, j \rangle}^C \cdot x_{\langle i, j \rangle} + (1 - c_{\langle i, j \rangle}^C) \cdot (1 - x_{\langle i, j \rangle}) \]
\[ \quad + \sum_{j \in M} c_j^A \cdot y_j + (1 - c_j^A) \cdot (1 - y_j) \]

subject to:
\[ x_{\langle i, j \rangle} \in \{0, 1\} \quad \forall \langle i, j \rangle \in P \]
\[ y_j \in \{0, 1\} \quad \forall y_j \in M \]

resolve all anaphors: \[ y_j \leq \sum_{i \in M[j]} x_{\langle i, j \rangle} \quad \forall j \in M \]
resolve only anaphors: \[ y_j \geq x_{\langle i, j \rangle} \quad \forall \langle i, j \rangle \in P \]
Joint Formulations (I)

Integrating DS model

\[
\begin{align*}
\text{maximize:} & \quad \sum_{\langle i, j \rangle \in P} c^{C}_{\langle i, j \rangle} \cdot x_{\langle i, j \rangle} + (1 - c^{C}_{\langle i, j \rangle}) \cdot (1 - x_{\langle i, j \rangle}) \\
& + \sum_{j \in M} c^{A}_{j} \cdot y_{j} + (1 - c^{A}_{j}) \cdot (1 - y_{j}) \\
\text{subject to:} & \quad x_{\langle i, j \rangle} \in \{0, 1\} \quad \forall \langle i, j \rangle \in P \\
& \quad y_{j} \in \{0, 1\} \quad \forall y_{j} \in M \\
& \quad y_{j} \leq \sum_{i \in M[j]} x_{\langle i, j \rangle} \quad \forall j \in M \\
& \quad y_{j} \geq x_{\langle i, j \rangle} \quad \forall \langle i, j \rangle \in P
\end{align*}
\]
Joint Formulations (I)

Integrating DS model

maximize:  \[ \sum_{\langle i, j \rangle \in P} c_{\langle i, j \rangle}^C \cdot x_{\langle i, j \rangle} + (1 - c_{\langle i, j \rangle}^C) \cdot (1 - x_{\langle i, j \rangle}) \]
\[ + \sum_{j \in M} c_A^j \cdot y_j + (1 - c_A^j) \cdot (1 - y_j) \]

subject to:
\[ x_{\langle i, j \rangle} \in \{0, 1\} \quad \forall \langle i, j \rangle \in P \]
\[ y_j \in \{0, 1\} \quad \forall y_j \in M \]
\[ y_j \leq \sum_{i \in M[j]} x_{\langle i, j \rangle} \quad \forall j \in M \]
\[ y_j \geq x_{\langle i, j \rangle} \quad \forall \langle i, j \rangle \in P \]
Joint Formulations (I)

Integrating DS model

maximize:
\[
\sum_{\langle i, j \rangle \in P} c_{\langle i, j \rangle}^C \cdot x_{\langle i, j \rangle} + (1 - c_{\langle i, j \rangle}^C) \cdot (1 - x_{\langle i, j \rangle})
\]
\[
+ \sum_{j \in M} c_j^A \cdot y_j + (1 - c_j^A) \cdot (1 - y_j)
\]

subject to:
\[
x_{\langle i, j \rangle} \in \{0, 1\} \quad \forall \langle i, j \rangle \in P
\]
\[
y_j \in \{0, 1\} \quad \forall y_j \in M
\]

resolve all anaphors:
\[
y_j \leq \sum_{i \in M[j]} x_{\langle i, j \rangle} \quad \forall j \in M
\]

resolve only anaphors:
\[
y_j \geq x_{\langle i, j \rangle} \quad \forall \langle i, j \rangle \in P
\]
Joint Formulations (II)

Integrating NEC model

- \( \mathcal{M} \): Set of mentions in document
- \( P \): Set of mention pairs: \( \langle i, j \rangle \in \mathcal{M} \times \mathcal{M}, i < j \)
- \( x_{\langle i, j \rangle} \): Coreference variables: 
  \[
  x_{\langle i, j \rangle} = \begin{cases} 
    1 & \text{if } i \text{ and } j \text{ corefer} \\
    0 & \text{otherwise}
  \end{cases}
  \]
- \( c^C \): Coreference Cost function: 
  \[
  c_{\langle i, j \rangle} = P_{Cl}(\text{COREF}|\langle i, j \rangle)
  \]
- \( z_{\langle i, t \rangle} \): NE type variables: 
  \[
  z_{\langle i, t \rangle} = \begin{cases} 
    1 & \text{if } i \text{ is of type } t \\
    0 & \text{otherwise}
  \end{cases}
  \]
- \( c^E \): NEC Cost function: 
  \[
  c_j^E = P_{ne}(t|i)
  \]
Joint Formulations (II)

Integrating NEC model

- $\mathcal{M}$: Set of mentions in document
- $P$: Set of mention pairs: $\langle i, j \rangle \in \mathcal{M} \times \mathcal{M}, i < j$
- $x_{\langle i, j \rangle}$: Coreference variables: $x_{\langle i, j \rangle} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ corefer} \\ 0 & \text{otherwise} \end{cases}$
- $c^C$: Coreference Cost function: $c_{\langle i, j \rangle} = P_{Cl}(\text{COREF}|\langle i, j \rangle)$
- $z_{\langle i, t \rangle}$: NE type variables: $z_{\langle i, t \rangle} = \begin{cases} 1 & \text{if } i \text{ is of type } t \\ 0 & \text{otherwise} \end{cases}$
- $c^E$: NEC Cost function: $c^E_j = P_{ne}(t|i)$
Joint Formulations (II)

Integrating NEC model

- \( \mathcal{M} \): Set of mentions in document
- \( P \): Set of mention pairs: \( \langle i, j \rangle \in \mathcal{M} \times \mathcal{M}, i < j \)
- \( x_{\langle i,j \rangle} \): Coreference variables: \( x_{\langle i,j \rangle} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ corefer} \\ 0 & \text{otherwise} \end{cases} \)
- \( c^C \): Coreference Cost function: \( c_{\langle i,j \rangle} = P_{ci}(\text{COREF}|\langle i,j \rangle) \)
- \( z_{\langle i,t \rangle} \): NE type variables: \( z_{\langle i,t \rangle} = \begin{cases} 1 & \text{if } i \text{ is of type } t \\ 0 & \text{otherwise} \end{cases} \)
- \( c^E \): NEC Cost function: \( c^E_j = P_{ne}(t|i) \)
Integrating NEC model

maximize: \[ \sum_{\langle i, j \rangle \in P} c^C_{\langle i, j \rangle} \cdot x_{\langle i, j \rangle} + (1 - c^C_{\langle i, j \rangle}) \cdot (1 - x_{\langle i, j \rangle}) \]
\[ + \sum_{\langle i, t \rangle \in M \times T} c^E_{\langle i, t \rangle} \cdot z_{\langle i, t \rangle} \]

subject to: \[ x_{\langle i, j \rangle} \in \{0, 1\} \quad \forall \langle i, j \rangle \in P \]
\[ z_{\langle i, t \rangle} \in \{0, 1\} \quad \forall \langle i, t \rangle \in M \times T \]

Coreferential mentions agree on type
\[ 1 - x_{\langle i, j \rangle} \geq |z_{\langle i, t \rangle} - z_{\langle j, t \rangle}| \quad \forall \langle i, j \rangle \in P, \forall t \in T \]
Joint Formulations (II)

Integrating NEC model

Maximize:
\[
\sum_{\langle i, j \rangle \in P} c^C_{\langle i, j \rangle} \cdot x_{\langle i, j \rangle} + (1 - c^C_{\langle i, j \rangle}) \cdot (1 - x_{\langle i, j \rangle})
\]
\[
+ \sum_{\langle i, t \rangle \in M \times T} c^E_{\langle i, t \rangle} \cdot z_{\langle i, t \rangle}
\]

Subject to:
\[
x_{\langle i, j \rangle} \in \{0, 1\} \quad \forall \langle i, j \rangle \in P
\]
\[
z_{\langle i, t \rangle} \in \{0, 1\} \quad \forall \langle i, t \rangle \in M \times T
\]

Coreferential mentions agree on type
\[
1 - x_{\langle i, j \rangle} \geq |z_{\langle i, t \rangle} - z_{\langle j, t \rangle}| \quad \forall \langle i, j \rangle \in P, \forall t \in T
\]
Joint Formulations (II)

Integrating NEC model

\[
\begin{align*}
\text{maximize:} & \quad \sum_{\langle i, j \rangle \in P} c_{\langle i, j \rangle}^C \cdot x_{\langle i, j \rangle} + (1 - c_{\langle i, j \rangle}^C) \cdot (1 - x_{\langle i, j \rangle}) \\
& \quad + \sum_{\langle i, t \rangle \in M \times T} c_{\langle i, t \rangle}^E \cdot z_{\langle i, t \rangle} \\
\text{subject to:} & \quad x_{\langle i, j \rangle} \in \{0, 1\} \quad \forall \langle i, j \rangle \in P \\
& \quad z_{\langle i, t \rangle} \in \{0, 1\} \quad \forall \langle i, t \rangle \in M \times T \\
& \quad \text{Coreferential mentions agree on type} \\
& \quad 1 - x_{\langle i, j \rangle} \geq |z_{\langle i, t \rangle} - z_{\langle j, t \rangle}| \quad \forall \langle i, j \rangle \in P, \forall t \in T
\end{align*}
\]
The “curse” of the pipeline

Example

Clinton told National Public Radio that his answers to questions about Lewinsky were constrained by Starr’s investigation. NPR reporter Mara Liasson asked Clinton "whether you had any conversations with her about her testimony, had any conversations at all."

NE Model

⟨“Clinton”, PER⟩ .83
⟨“Nat’l Public Radio”, ORG⟩ .61
⟨“NPR”, GPE⟩ .59
... ... 

Coreference Model

⟨“Clinton”, “Nat’l Public Radio”⟩ .02
⟨“Clinton”, “his”⟩ .82
⟨“Clinton”, “Lewinsky”⟩ .01
⟨“Nat’l Public Radio”, “NPR”⟩ .92
... ... 

P. Denis

Balancing multiple models for coref. resolution using ILP
Effects of combining the three models

- By mutually constraining: (i) coreference with anaphoricity, and (ii) coreference with NE, we get “free” anaphoricity-NE constraints (there is no anaphor of type $t$ if not preceded by another mention of type $t$)

- The addition of the NE model gives a *global* model: “chaining” effects from coreference to NE back to coreference

- But anaphoricity constraints don’t get propagated all the way: e.g. $\langle i, k \rangle = 1$, $\langle j, k \rangle = 1$ but $y_j = 0$
### Experiment #3: Simple ILP vs. Joint formulations

<table>
<thead>
<tr>
<th>System</th>
<th>MUC</th>
<th></th>
<th>B³</th>
<th></th>
<th>CEAF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R</td>
<td>P</td>
<td>F</td>
<td>R</td>
</tr>
<tr>
<td>COREF&lt;sub&gt;base&lt;/sub&gt;</td>
<td>60.8</td>
<td>72.6</td>
<td>66.2</td>
<td>62.4</td>
<td>77.7</td>
</tr>
<tr>
<td>ILP&lt;sub&gt;c&lt;/sub&gt;</td>
<td>70.3</td>
<td>72.7</td>
<td>71.5</td>
<td>73.2</td>
<td>63.7</td>
</tr>
<tr>
<td>ILP&lt;sub&gt;c,a&lt;/sub&gt;</td>
<td>73.2</td>
<td>73.4</td>
<td>73.3</td>
<td>75.3</td>
<td>62.0</td>
</tr>
<tr>
<td>ILP&lt;sub&gt;c,e&lt;/sub&gt;</td>
<td>66.2</td>
<td>75.0</td>
<td>70.4</td>
<td>69.6</td>
<td>71.2</td>
</tr>
<tr>
<td>ILP&lt;sub&gt;c,a,e&lt;/sub&gt;</td>
<td>69.6</td>
<td>75.4</td>
<td>72.4</td>
<td>72.2</td>
<td>69.7</td>
</tr>
</tbody>
</table>

Accuracy of NE classifier also went up, from 79.5 to 80.5!
## Experiment #3: Simple ILP vs. Joint formulations

<table>
<thead>
<tr>
<th>System</th>
<th>MUC</th>
<th>B³</th>
<th>CEAF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F</td>
</tr>
<tr>
<td>COREF&lt;sub&gt;base&lt;/sub&gt;</td>
<td>60.8</td>
<td>72.6</td>
<td>66.2</td>
</tr>
<tr>
<td>ILP&lt;sub&gt;c&lt;/sub&gt;</td>
<td>70.3</td>
<td>72.7</td>
<td>71.5</td>
</tr>
<tr>
<td>ILP&lt;sub&gt;c,a&lt;/sub&gt;</td>
<td>73.2</td>
<td>73.4</td>
<td>73.3</td>
</tr>
<tr>
<td>ILP&lt;sub&gt;c,e&lt;/sub&gt;</td>
<td>66.2</td>
<td>75.0</td>
<td>70.4</td>
</tr>
<tr>
<td>ILP&lt;sub&gt;c,a,e&lt;/sub&gt;</td>
<td>69.6</td>
<td>75.4</td>
<td>72.4</td>
</tr>
</tbody>
</table>

Accuracy of NE classifier also went up, from 79.5 to 80.5!
Rest of the talk

1. Base Models
   - Coreference classifier
   - Discourse status classifier
   - Named entity classifier
   - Data and Evaluation
   - Pipeline experiments

2. ILP formulations
   - Simple formulation
   - Joint formulations
   - Adding global constraints

3. Summary and future work
Triangular constraints

ILP makes it easy to encode global constraints between coreference assignments:

- **Transitivity**

- **Euclideanity**

- **Anti-Euclideanity**
ILP makes it easy to encode global constraints between coreference assignments:

- Transitivity

  ![Transitivity Diagram](image)

- Euclideanity

  ![Euclideanity Diagram](image)

- Anti-Euclideanity
Triangular constraints (cont’d)

- **Transitivity**: \( x_{i,k} \geq x_{i,j} + x_{j,k} - 1, \forall \langle i, j, k \rangle \in M_{i,j,k} \)
- **Euclideanity**: \( x_{i,j} \geq x_{i,k} + x_{j,k} - 1, \forall \langle i, j, k \rangle \in M_{i,j,k} \)
- **Anti-Euclideanity**: \( x_{i,j} \geq x_{i,k} + x_{j,k} - 1, \forall \langle i, j, k \rangle \in M_{i,j,k} \)
Experiment #4: Simple ILP vs. Joint formulations with trans. constraints

<table>
<thead>
<tr>
<th>System</th>
<th>MUC</th>
<th>B³</th>
<th>CEAF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F</td>
</tr>
<tr>
<td><strong>COREF</strong>&lt;sub&gt;base&lt;/sub&gt;</td>
<td>60.8</td>
<td>72.6</td>
<td><strong>66.2</strong></td>
</tr>
<tr>
<td><strong>ILP</strong>&lt;sub&gt;c&lt;/sub&gt;</td>
<td>70.3</td>
<td>72.7</td>
<td><strong>71.5</strong></td>
</tr>
<tr>
<td><strong>ILP</strong>&lt;sub&gt;c,a&lt;/sub&gt;</td>
<td>73.2</td>
<td>73.4</td>
<td><strong>73.3</strong></td>
</tr>
<tr>
<td><strong>ILP</strong>&lt;sub&gt;c,e&lt;/sub&gt;</td>
<td>66.2</td>
<td>75.0</td>
<td><strong>70.4</strong></td>
</tr>
<tr>
<td><strong>ILP</strong>&lt;sub&gt;c,a,e&lt;/sub&gt;</td>
<td>69.6</td>
<td>75.4</td>
<td><strong>72.4</strong></td>
</tr>
<tr>
<td><strong>ILP</strong>&lt;sub&gt;c,a,e</td>
<td>trans&lt;/sub&gt;</td>
<td>63.7</td>
<td>77.8</td>
</tr>
</tbody>
</table>
Rest of the talk

1. Base Models
   - Coreference classifier
   - Discourse status classifier
   - Named entity classifier
   - Data and Evaluation
   - Pipeline experiments

2. ILP formulations
   - Simple formulation
   - Joint formulations
   - Adding global constraints

3. Summary and future work
Summary

- ILP provides an elegant way to model dependencies between three models and encode global constraints.
- Our ILP formulation delivers significant $f$-score gains over the base coreference model on the 3 coreference metrics.
- Additional gains are obtained through the use of anti-Euclidean constraints on $B^3$ and CEAF.
- MUC metric is unreliable due to its preference for larger entities.
- Other possible global constraints: number of anaphors/entities per document, number of links per document.
- End-to-end coreference: mention detection into the mix?