Modeling Unrestricted Coreference in OntoNotes

CoNLL-2011 Shared Task

Sameer S Pradhan[1] Lance Ramshaw[1] Mitchell Marcus[2]
Martha Palmer[3] Ralph Weischedel[1] Nianwen Xue[4]

[1]BBN Technologies, Cambridge, MA
[2]University of Pennsylvania, Philadelphia, PA
[3]University of Colorado, Boulder, CO
[4]Brandeis University, Waltham, MA

CoNLL Shared Task: Pushing the State of the Art

- This is the **12^{th} year** of the CoNLL shared task

Year	Task
2000	Base Phrase chunking
2001	Clause identification
2002, 2003	Named Entity recognition
2004, 2005	Semantic Role Labeling
2006, 2007	Syntactic dependency parsing
2008, 2009	Syntactic and semantic dependency parsing
2010	Hedge detection
2011	Coreference resolution

CoNLL Shared Task: Pushing the State of the Art

- This is the **12th year** of the CoNLL shared task

Year	Task
2000	Base Phrase chunking
2001	Clause identification
2002, 2003	Named Entity recognition
2004, 2005	Semantic Role Labeling
2006, 2007	Syntactic dependency parsing
2008, 2009	Syntactic and semantic dependency parsing
2010	Hedge detection
2011	Coreference resolution

CoNLL Shared Task: Pushing the State of the Art

- This is the **12th year** of the CoNLL shared task

Year	Task
2000	Base Phrase chunking
2001	Clause identification
2002, 2003	Named Entity recognition
2004, 2005	Semantic Role Labeling
2006, 2007	Syntactic dependency parsing
2008, 2009	Syntactic and semantic dependency parsing
2010	Hedge detection
2011	Coreference resolution

CoNLL Shared Task: Pushing the State of the Art

- This is the **12^{th} year** of the CoNLL shared task

Year	Task
2000	Base Phrase chunking
2001	Clause identification
2002, 2003	Named Entity recognition
2004, 2005	Semantic Role Labeling
2006, 2007	Syntactic dependency parsing
2008, 2009	Syntactic and semantic dependency parsing
2010	Hedge detection
2011	Coreference resolution

Why Coreference?

- Wasn't tackled before as a CoNLL Shared Task
- Higher level task which could benefit from other layers
- Not much coreference data available before for *unrestricted* types of entities and events
- No standard evaluation set
- OntoNotes + CoNLL = Standard benchmark

Why Coreference?

- Wasn't tackled before as a CoNLL Shared Task
- Higher level task which could benefit from other layers
- Not much coreference data available before for *unrestricted* types of entities and events
- No standard evaluation set
- OntoNotes + CoNLL = Standard benchmark

Why Coreference?

- Wasn't tackled before as a CoNLL Shared Task
- Higher level task which could benefit from other layers
- Not much coreference data available before for *unrestricted* types of entities and events
- No standard evaluation set
- OntoNotes + CoNLL = Standard benchmark

Why Coreference?

- Wasn't tackled before as a CoNLL Shared Task
- Higher level task which could benefit from other layers
- Not much coreference data available before for *unrestricted* types of entities and events
- No standard evaluation set
- OntoNotes + CoNLL = Standard benchmark

Why Coreference?

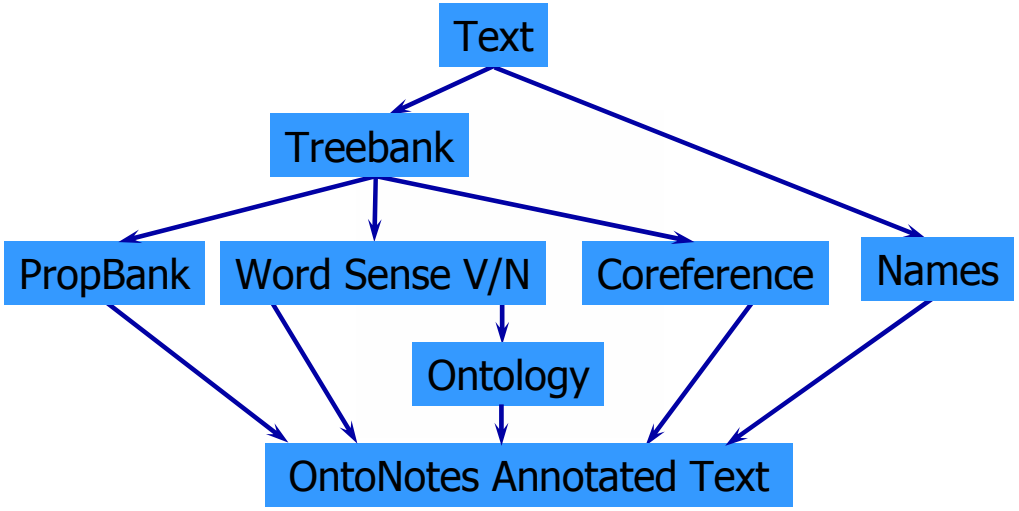
- Wasn't tackled before as a CoNLL Shared Task
- Higher level task which could benefit from other layers
- Not much coreference data available before for *unrestricted* types of entities and events
- No standard evaluation set
- OntoNotes + CoNLL = Standard benchmark

Why Coreference?

- Wasn't tackled before as a CoNLL Shared Task
- Higher level task which could benefit from other layers
- Not much coreference data available before for *unrestricted* types of entities and events
- No standard evaluation set
- OntoNotes + CoNLL = Standard benchmark

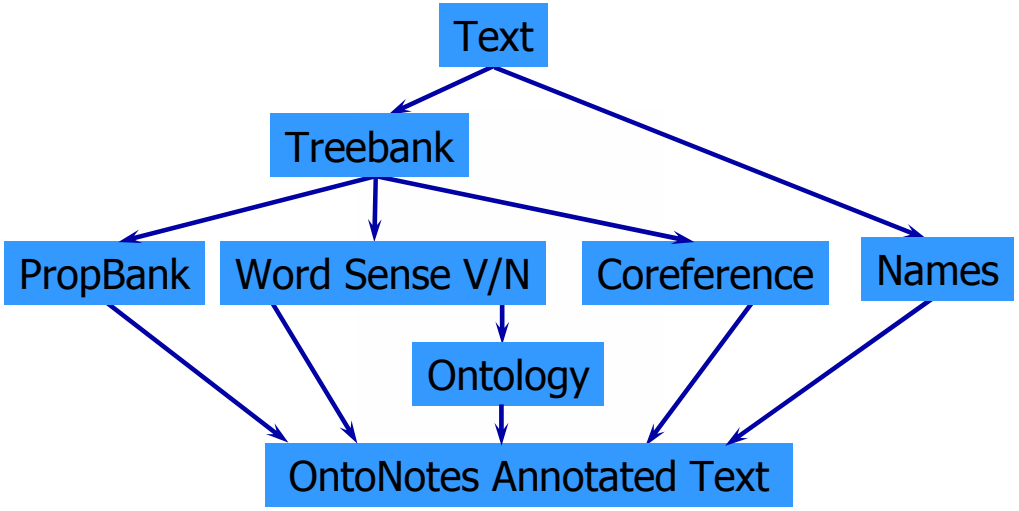
OntoNotes: Large Annotated Corpus

- Multiple layers of annotation
  - Syntax
  - Propositions
  - Word sense
  - Coreference
  - Names
- Multiple Languages
  - English ($\sim$ 1.3 MW)
  - Chinese ($\sim$ 1 MW)
  - Arabic ($\sim$ 3 KW)
- Multiple Genres
  - Newswire
  - Broadcast News
  - Broadcast Conversation
  - Web Newsgroups and Blogs
  - Telephone Conversation
- High Inter-Annotator Agreement



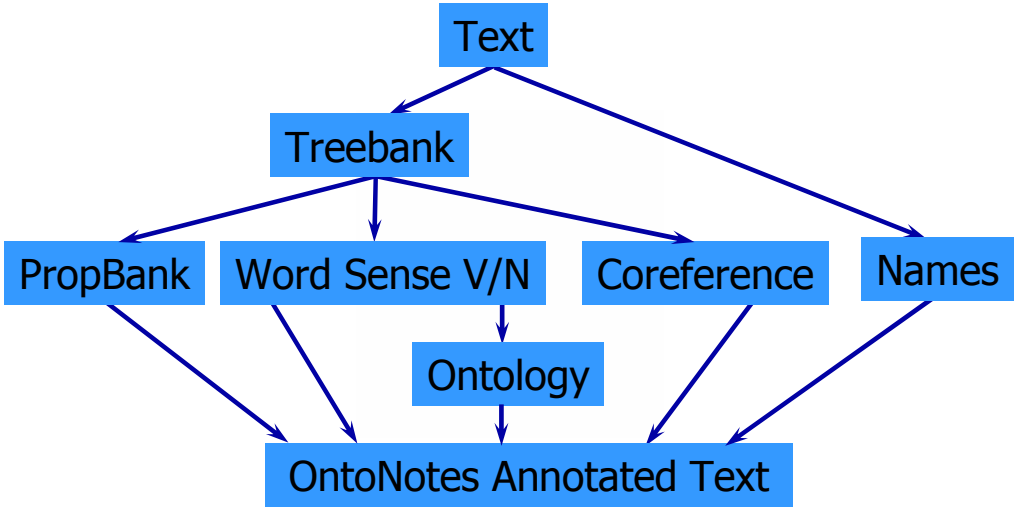
OntoNotes: Large Annotated Corpus

- **Multiple layers of annotation**
  - Syntax
  - Propositions
  - Word sense
  - Coreference
  - Names
- **Multiple Languages**
  - English ($\sim$ 1.3 MW)
  - Chinese ($\sim$ 1 MW)
  - Arabic ($\sim$ 3 KW)
- **Multiple Genres**
  - Newswire
  - Broadcast News
  - Broadcast Conversation
  - Web Newsgroups and Blogs
  - Telephone Conversation
- **High Inter-Annotator Agreement**



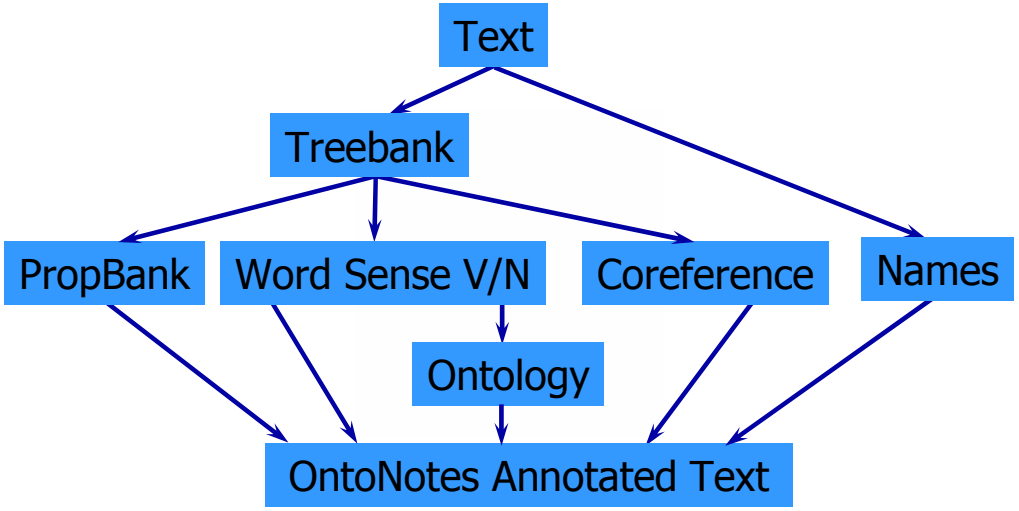
OntoNotes: Large Annotated Corpus

- **Multiple layers of annotation**
  - Syntax
  - Propositions
  - Word sense
  - Coreference
  - Names
- **Multiple Languages**
  - English ($\sim$ 1.3 MW)
  - Chinese ($\sim$ 1 MW)
  - Arabic ($\sim$ 3 KW)
- **Multiple Genres**
  - Newswire
  - Broadcast News
  - Broadcast Conversation
  - Web Newsgroups and Blogs
  - Telephone Conversation
- **High Inter-Annotator Agreement**



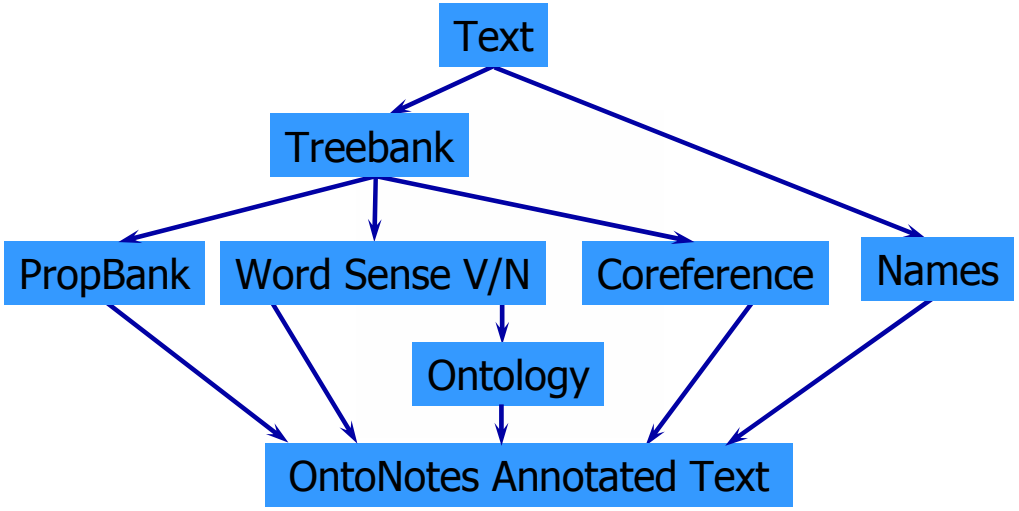
OntoNotes: Large Annotated Corpus

- **Multiple layers of annotation**
  - Syntax
  - Propositions
  - Word sense
  - Coreference
  - Names
- **Multiple Languages**
  - English ($\sim$ 1.3 MW)
  - Chinese ($\sim$ 1 MW)
  - Arabic ($\sim$ 3 KW)
- **Multiple Genres**
  - Newswire
  - Broadcast News
  - Broadcast Conversation
  - Web Newsgroups and Blogs
  - Telephone Conversation
- **High Inter-Annotator Agreement**



OntoNotes: Large Annotated Corpus

- **Multiple layers of annotation**
  - Syntax
  - Propositions
  - Word sense
  - Coreference
  - Names
- **Multiple Languages**
  - English ($\sim$ 1.3 MW)
  - Chinese ($\sim$ 1 MW)
  - Arabic ($\sim$ 3 KW)
- **Multiple Genres**
  - Newswire
  - Broadcast News
  - Broadcast Conversation
  - Web Newsgroups and Blogs
  - Telephone Conversation
- **High Inter-Annotator Agreement**



Characteristics of OntoNotes Coreference

- **Much more data** linking all entity and event types

MUC	60K words over 120 documents
OntoNotes	1.3M words over 2K documents

- Spans **five** genres
- **Both** Entities and Events
- **No singletons** – only multi-mention entities annotated
- Two types of coreference
  - **IDENT**ity
  - **APPOS**itive
- **No** Copular constructions
- **No** Generics, or underspecified mentions
- Mentions tagged on **Treebank NPs, verbs** and names (∼2% exception)

Characteristics of OntoNotes Coreference

- **Much more data** linking all entity and event types

MUC	60K words over 120 documents
OntoNotes	1.3M words over 2K documents

- Spans **five** genres
- **Both** Entities and Events
- **No singletons** – only multi-mention entities annotated
- Two types of coreference
  - **IDENT**ity
  - **APPOS**itive
- **No** Copular constructions
- **No** Generics, or underspecified mentions
- Mentions tagged on **Treebank NPs, verbs** and names (∼2% exception)

Characteristics of OntoNotes Coreference

- **Much more data** linking all entity and event types

MUC	60K words over 120 documents
OntoNotes	1.3M words over 2K documents

- Spans **five** genres
- **Both** Entities and Events
- **No singletons** – only multi-mention entities annotated
- Two types of coreference
  - **IDENT**ity
  - **APPOS**itive
- **No** Copular constructions
- **No** Generics, or underspecified mentions
- Mentions tagged on **Treebank NPs, verbs** and names (∼2% exception)

Characteristics of OntoNotes Coreference

- **Much more data** linking all entity and event types

MUC	60K words over 120 documents
OntoNotes	1.3M words over 2K documents

- Spans **five** genres
- **Both** Entities and Events
- **No singletons** – only multi-mention entities annotated
- Two types of coreference
  - **IDENT**ity
  - **APPOS**itive
- **No** Copular constructions
- **No** Generics, or underspecified mentions
- Mentions tagged on **Treebank NPs, verbs** and names (∼2% exception)

Characteristics of OntoNotes Coreference

- **Much more data** linking all entity and event types

MUC	60K words over 120 documents
OntoNotes	1.3M words over 2K documents

- Spans **five** genres
- **Both** Entities and Events
- **No singletons** – only multi-mention entities annotated
- Two types of coreference
  - **IDENT**ity
  - **APPOS**itive
- **No** Copular constructions
- **No** Generics, or underspecified mentions
- Mentions tagged on **Treebank NPs, verbs** and names (∼2% exception)

Characteristics of OntoNotes Coreference

- **Much more data** linking all entity and event types

MUC	60K words over 120 documents
OntoNotes	1.3M words over 2K documents

- Spans **five** genres
- **Both** Entities and Events
- **No singletons** – only multi-mention entities annotated
- Two types of coreference
  - **IDENT**ity
  - **APPOS**itive
- **No** Copular constructions
- **No** Generics, or underspecified mentions
- Mentions tagged on **Treebank NPs, verbs** and names (∼2% exception)

Characteristics of OntoNotes Coreference

- **Much more data** linking all entity and event types

MUC	60K words over 120 documents
OntoNotes	1.3M words over 2K documents

- Spans **five** genres
- **Both** Entities and Events
- **No singletons** – only multi-mention entities annotated
- Two types of coreference
  - **IDENT**ity
  - **APPOS**itive
- **No** Copular constructions
- **No** Generics, or underspecified mentions
- Mentions tagged on **Treebank NPs, verbs** and names (∼2% exception)

Characteristics of OntoNotes Coreference

- **Much more data** linking all entity and event types

MUC	60K words over 120 documents
OntoNotes	1.3M words over 2K documents

- Spans **five** genres
- **Both** Entities and Events
- **No singletons** – only multi-mention entities annotated
- Two types of coreference
  - **IDENT**ity
  - **APPOS**itive
- **No** Copular constructions
- **No** Generics, or underspecified mentions
- Mentions tagged on **Treebank NPs, verbs** and names (∼2% exception)

Characteristics of OntoNotes Coreference

- **Much more data** linking all entity and event types

MUC	60K words over 120 documents
OntoNotes	1.3M words over 2K documents

- Spans **five** genres
- **Both** Entities and Events
- **No singletons** – only multi-mention entities annotated
- Two types of coreference
  - **IDENT**ity
  - **APPOS**itive
- **No** Copular constructions
- **No** Generics, or underspecified mentions
- Mentions tagged on **Treebank NPs, verbs** and names (∼2% exception)

Characteristics of OntoNotes Coreference

- **Much more data** linking all entity and event types

MUC	60K words over 120 documents
OntoNotes	1.3M words over 2K documents

- Spans **five** genres
- **Both** Entities and Events
- **No singletons** – only multi-mention entities annotated
- Two types of coreference
  - **IDENT**ity
  - **APPOS**itive
- **No** Copular constructions
- **No** Generics, or underspecified mentions
- Mentions tagged on **Treebank NPs, verbs** and names (∼2% exception)

Characteristics of OntoNotes Coreference

- **Much more data** linking all entity and event types

MUC	60K words over 120 documents
OntoNotes	1.3M words over 2K documents

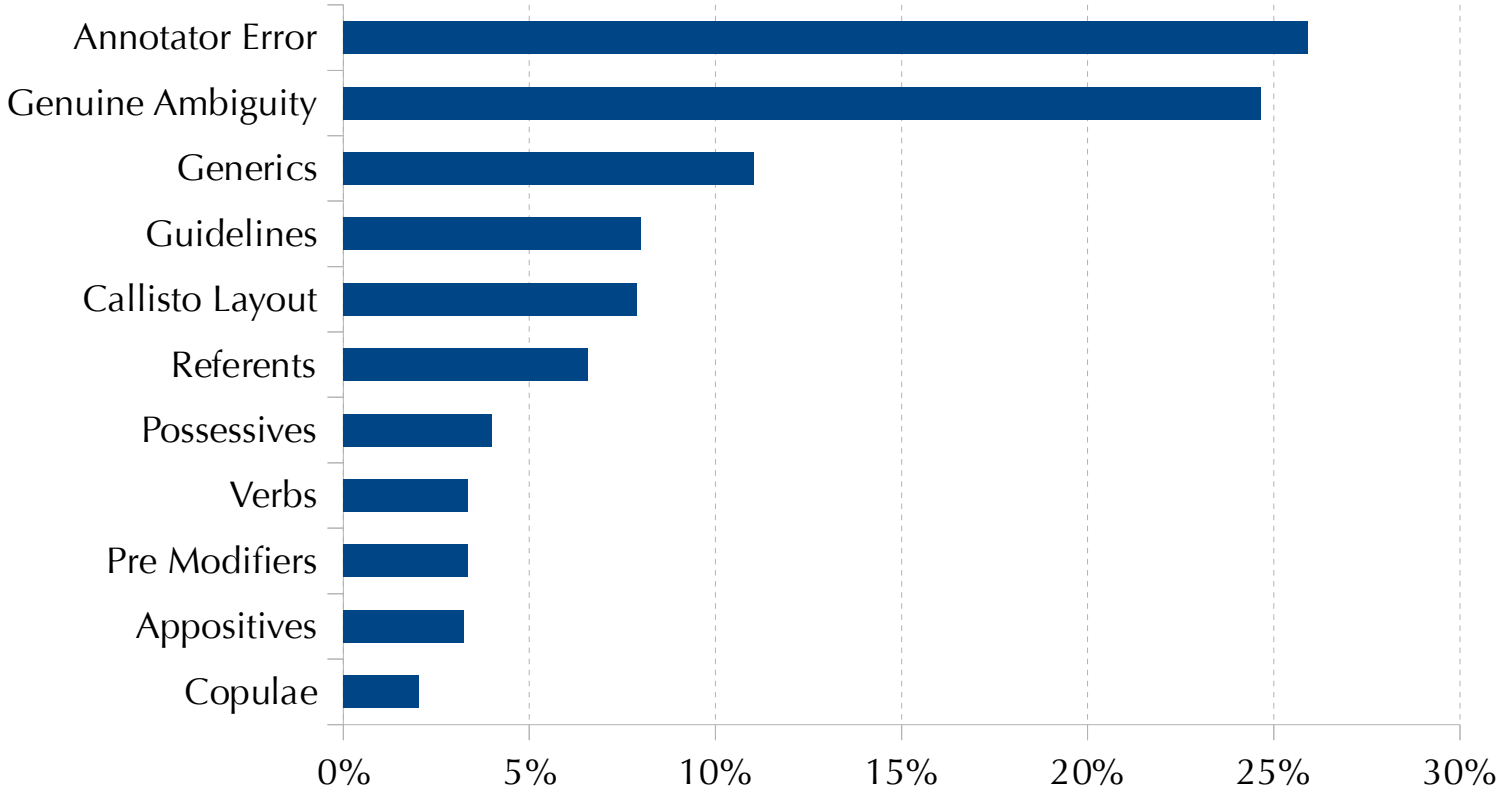
- Spans **five** genres
- **Both** Entities and Events
- **No singletons** – only multi-mention entities annotated
- Two types of coreference
  - **IDENT**ity
  - **APPOS**itive
- **No** Copular constructions
- **No** Generics, or underspecified mentions
- Mentions tagged on **Treebank NPs, verbs** and names (∼2% exception)

Inter-Annotator Agreement

Genre	ANN1-ANN2	ANN1-ADJ	ANN2-ADJ
Newswire	80.9	85.2	88.3
Broadcast News	78.6	83.5	89.4
Broadcast Conversation	86.7	91.6	93.7
Magazine	78.4	83.2	88.8
Web	85.9	92.2	91.2

Disagreement Distribution

- Disagreements over a sample of 11K words



Possible Evaluation Parameters

- **Scope**
  - Identify and cluster all anaphoric mentions
  - Cluster all anaphoric mentions given possible mention boundaries
  - Cluster all anaphoric mentions given correct mentions
- **Specificity**
  - Exact phrase match
  - Weighted/unweighted head word match
- **Quality of Layers**
  - Gold standard
  - Predicted layers
- **Metric**
  - No silver-bullet
  - Various proposed over the years
    1. MUC [Vilain et al., 1995] (link based)
    2. B-CUBED [Bagga and Baldwin, 1998] (mention based)
    3. CEAF$_{m/e}$ [Luo, 2005] (entity based)
    4. BLANC [Recasens and Hovy, 2011] (Rand-index based)
- **External Resources**

Possible Evaluation Parameters

- **Scope**
  - Identify and cluster all anaphoric mentions
  - Cluster all anaphoric mentions given possible mention boundaries
  - Cluster all anaphoric mentions given correct mentions
- **Specificity**
  - Exact phrase match
  - Weighted/unweighted head word match
- **Quality of Layers**
  - Gold standard
  - Predicted layers
- **Metric**
  - No silver-bullet
  - Various proposed over the years
    1. MUC [Vilain et al., 1995] (link based)
    2. B-CUBED [Bagga and Baldwin, 1998] (mention based)
    3. CEAF$_{m/e}$ [Luo, 2005] (entity based)
    4. BLANC [Recasens and Hovy, 2011] (Rand-index based)
- **External Resources**

Possible Evaluation Parameters

- **Scope**
  - Identify and cluster all anaphoric mentions
  - Cluster all anaphoric mentions given possible mention boundaries
  - Cluster all anaphoric mentions given correct mentions
- **Specificity**
  - Exact phrase match
  - Weighted/unweighted head word match
- **Quality of Layers**
  - Gold standard
  - Predicted layers
- **Metric**
  - No silver-bullet
  - Various proposed over the years
    1. MUC [Vilain et al., 1995] (link based)
    2. B-CUBED [Bagga and Baldwin, 1998] (mention based)
    3. $\text{CEAF}_{m/e}$ [Luo, 2005] (entity based)
    4. BLANC [Recasens and Hovy, 2011] (Rand-index based)
- **External Resources**

Possible Evaluation Parameters

- **Scope**
  - Identify and cluster all anaphoric mentions
  - Cluster all anaphoric mentions given possible mention boundaries
  - Cluster all anaphoric mentions given correct mentions
- **Specificity**
  - Exact phrase match
  - Weighted/unweighted head word match
- **Quality of Layers**
  - Gold standard
  - Predicted layers
- **Metric**
  - No silver-bullet
  - Various proposed over the years
    1. MUC [Vilain et al., 1995] (link based)
    2. B-CUBED [Bagga and Baldwin, 1998] (mention based)
    3. CEAF$_{m/e}$ [Luo, 2005] (entity based)
    4. BLANC [Recasens and Hovy, 2011] (Rand-index based)
- **External Resources**

Possible Evaluation Parameters

- **Scope**
  - Identify and cluster all anaphoric mentions
  - Cluster all anaphoric mentions given possible mention boundaries
  - Cluster all anaphoric mentions given correct mentions
- **Specificity**
  - Exact phrase match
  - Weighted/unweighted head word match
- **Quality of Layers**
  - Gold standard
  - Predicted layers
- **Metric**
  - No silver-bullet
  - Various proposed over the years
    1. MUC [Vilain et al., 1995] (link based)
    2. B-CUBED [Bagga and Baldwin, 1998] (mention based)
    3. $\text{CEAF}_{m/e}$ [Luo, 2005] (entity based)
    4. BLANC [Recasens and Hovy, 2011] (Rand-index based)
- **External Resources**

Possible Evaluation Parameters

- **Scope**
  - Identify and cluster all anaphoric mentions
  - Cluster all anaphoric mentions given possible mention boundaries
  - Cluster all anaphoric mentions given correct mentions
- **Specificity**
  - Exact phrase match
  - Weighted/unweighted head word match
- **Quality of Layers**
  - Gold standard
  - Predicted layers
- **Metric**
  - No silver-bullet
  - Various proposed over the years
    1. MUC [Vilain et al., 1995] (link based)
    2. B-CUBED [Bagga and Baldwin, 1998] (mention based)
    3. $\text{CEAF}_{m/e}$ [Luo, 2005] (entity based)
    4. BLANC [Recasens and Hovy, 2011] (Rand-index based)
- **External Resources**

Possible Evaluation Parameters

- **Scope**
  - Identify and cluster all anaphoric mentions
  - Cluster all anaphoric mentions given possible mention boundaries
  - Cluster all anaphoric mentions given correct mentions
- **Specificity**
  - Exact phrase match
  - Weighted/unweighted head word match
- **Quality of Layers**
  - Gold standard
  - Predicted layers
- **Metric**
  - No silver-bullet
  - Various proposed over the years
    1. MUC [Vilain et al., 1995] (link based)
    2. B-CUBED [Bagga and Baldwin, 1998] (mention based)
    3. $\text{CEAF}_{m/e}$ [Luo, 2005] (entity based)
    4. BLANC [Recasens and Hovy, 2011] (Rand-index based)
- **External Resources**

Possible Evaluation Parameters

- **Scope**
  - Identify and cluster all anaphoric mentions
  - Cluster all anaphoric mentions given possible mention boundaries
  - Cluster all anaphoric mentions given correct mentions
- **Specificity**
  - Exact phrase match
  - Weighted/unweighted head word match
- **Quality of Layers**
  - Gold standard
  - Predicted layers
- **Metric**
  - No silver-bullet
  - Various proposed over the years
    1. MUC [Vilain et al., 1995] (link based)
    2. B-CUBED [Bagga and Baldwin, 1998] (mention based)
    3. $\text{CEAF}_{m/e}$ [Luo, 2005] (entity based)
    4. BLANC [Recasens and Hovy, 2011] (Rand-index based)
- **External Resources**

Possible Evaluation Parameters

- **Scope**
  - Identify and cluster all anaphoric mentions
  - Cluster all anaphoric mentions given possible mention boundaries
  - Cluster all anaphoric mentions given correct mentions
- **Specificity**
  - Exact phrase match
  - Weighted/unweighted head word match
- **Quality of Layers**
  - Gold standard
  - Predicted layers
- **Metric**
  - No silver-bullet
  - Various proposed over the years
    1. MUC [Vilain et al., 1995] (link based)
    2. B-CUBED [Bagga and Baldwin, 1998] (mention based)
    3. $\text{CEAF}_{m/e}$ [Luo, 2005] (entity based)
    4. BLANC [Recasens and Hovy, 2011] (Rand-index based)
- **External Resources**

Possible Evaluation Parameters

- **Scope**
  - Identify and cluster all anaphoric mentions
  - Cluster all anaphoric mentions given possible mention boundaries
  - Cluster all anaphoric mentions given correct mentions
- **Specificity**
  - Exact phrase match
  - Weighted/unweighted head word match
- **Quality of Layers**
  - Gold standard
  - Predicted layers
- **Metric**
  - No silver-bullet
  - Various proposed over the years
    1. MUC [Vilain et al., 1995] (link based)
    2. B-CUBED [Bagga and Baldwin, 1998] (mention based)
    3. $\text{CEAF}_{m/e}$ [Luo, 2005] (entity based)
    4. BLANC [Recasens and Hovy, 2011] (Rand-index based)
- **External Resources**

Possible Evaluation Parameters

- **Scope**
  - Identify and cluster all anaphoric mentions
  - Cluster all anaphoric mentions given possible mention boundaries
  - Cluster all anaphoric mentions given correct mentions
- **Specificity**
  - Exact phrase match
  - Weighted/unweighted head word match
- **Quality of Layers**
  - Gold standard
  - Predicted layers
- **Metric**
  - No silver-bullet
  - Various proposed over the years
    1. MUC [Vilain et al., 1995] (link based)
    2. B-CUBED [Bagga and Baldwin, 1998] (mention based)
    3. $\text{CEAF}_{m/e}$ [Luo, 2005] (entity based)
    4. BLANC [Recasens and Hovy, 2011] (Rand-index based)
- **External Resources**

Official Evaluation Parameters

- **Scope**
  - Identify and cluster all anaphoric mentions in English, that represent identity coreference (IDENT)
- **Specificity**
  - Exact phrase match
- **Quality of Layers**
  - Predicted layers
- **Metric**
  - Compute all metrics
  - Winning system determined by $\frac{MUC + B-CUBED + CEAF_e}{3}$
- **External Resources**
  - **Closed Track**
    - WordNet [Fellbaum, 1998]
    - Number and Gender table [Bergsma and Lin, 2006]
  - **Open Track**
    - Everything used in Closed Track, plus other resources, such as Wikipedia

Official Evaluation Parameters

- **Scope**
  - Identify and cluster all anaphoric mentions in English, that represent identity coreference (IDENT)
- **Specificity**
  - Exact phrase match
- **Quality of Layers**
  - Predicted layers
- **Metric**
  - Compute all metrics
  - Winning system determined by $\frac{MUC + B-CUBED + CEAF_e}{3}$
- **External Resources**
  - **Closed Track**
    - WordNet [Fellbaum, 1998]
    - Number and Gender table [Bergsma and Lin, 2006]
  - **Open Track**
    - Everything used in Closed Track, plus other resources, such as Wikipedia

Official Evaluation Parameters

- **Scope**
  - Identify and cluster all anaphoric mentions in English, that represent identity coreference (IDENT)
- **Specificity**
  - Exact phrase match
- **Quality of Layers**
  - Predicted layers
- **Metric**
  - Compute all metrics
  - Winning system determined by $\frac{MUC + B-CUBED + CEAF_e}{3}$
- **External Resources**
  - **Closed Track**
    - WordNet [Fellbaum, 1998]
    - Number and Gender table [Bergsma and Lin, 2006]
  - **Open Track**
    - Everything used in Closed Track, plus other resources, such as Wikipedia

Official Evaluation Parameters

- **Scope**
  - Identify and cluster all anaphoric mentions in English, that represent identity coreference (IDENT)
- **Specificity**
  - Exact phrase match
- **Quality of Layers**
  - Predicted layers
- **Metric**
  - Compute all metrics
  - Winning system determined by $\frac{MUC + B-CUBED + CEAF_e}{3}$
- **External Resources**
  - **Closed Track**
    - WordNet [Fellbaum, 1998]
    - Number and Gender table [Bergsma and Lin, 2006]
  - **Open Track**
    - Everything used in Closed Track, plus other resources, such as Wikipedia

Official Evaluation Parameters

- **Scope**
  - Identify and cluster all anaphoric mentions in English, that represent identity coreference (IDENT)
- **Specificity**
  - Exact phrase match
- **Quality of Layers**
  - Predicted layers
- **Metric**
  - Compute all metrics
  - Winning system determined by $\frac{MUC + B-CUBED + CEAF_e}{3}$
- **External Resources**
  - **Closed Track**
    - WordNet [Fellbaum, 1998]
    - Number and Gender table [Bergsma and Lin, 2006]
  - **Open Track**
    - Everything used in Closed Track, plus other resources, such as Wikipedia

Official Evaluation Parameters

- **Scope**
  - Identify and cluster all anaphoric mentions in English, that represent identity coreference (IDENT)
- **Specificity**
  - Exact phrase match
- **Quality of Layers**
  - Predicted layers
- **Metric**
  - Compute all metrics
  - Winning system determined by $\frac{MUC + B-CUBED + CEAF_e}{3}$
- **External Resources**
  - **Closed Track**
    - WordNet [Fellbaum, 1998]
    - Number and Gender table [Bergsma and Lin, 2006]
  - **Open Track**
    - Everything used in Closed Track, plus other resources, such as Wikipedia

Official Evaluation Parameters

- **Scope**
  - Identify and cluster all anaphoric mentions in English, that represent identity coreference (IDENT)
- **Specificity**
  - Exact phrase match
- **Quality of Layers**
  - Predicted layers
- **Metric**
  - Compute all metrics
  - Winning system determined by $\frac{MUC + B-CUBED + CEAF_e}{3}$
- **External Resources**
  - **Closed Track**
    - WordNet [Fellbaum, 1998]
    - Number and Gender table [Bergsma and Lin, 2006]
  - **Open Track**
    - Everything used in Closed Track, plus other resources, such as Wikipedia

Official Evaluation Parameters

- **Scope**
  - Identify and cluster all anaphoric mentions in English, that represent identity coreference (IDENT)
- **Specificity**
  - Exact phrase match
- **Quality of Layers**
  - Predicted layers
- **Metric**
  - Compute all metrics
  - Winning system determined by $\frac{MUC + B-CUBED + CEAF_e}{3}$
- **External Resources**
  - **Closed Track**
    - WordNet [Fellbaum, 1998]
    - Number and Gender table [Bergsma and Lin, 2006]
  - **Open Track**
    - Everything used in Closed Track, plus other resources, such as Wikipedia

Data Sample

```
#begin document (nw/wsj/07/wsj_0771); part 000
...
...
nw/wsj/07/wsj_0771 0   0         ``   ``  (TOP(S(S*     -    -  -   -         *       *     (ARG1*  *       -
nw/wsj/07/wsj_0771 0   1 Vandenberg  NNP       (NP*     -    -  -   -  (PERSON)  (ARG1*          *  *  (8|(0)
nw/wsj/07/wsj_0771 0   2        and   CC         *      -    -  -   -         *       *          *  *       -
nw/wsj/07/wsj_0771 0   3    Rayburn  NNP         *)     -    -  -   -  (PERSON)      *)         *  * (23)|8)
nw/wsj/07/wsj_0771 0   4        are  VBP       (VP*    be   01  1   -         *    (V*)         *  *       -
nw/wsj/07/wsj_0771 0   5     heroes  NNS   (NP(NP*)     -    -  -   -         *  (ARG2*         *  *       -
nw/wsj/07/wsj_0771 0   6         of   IN       (PP*     -    -  -   -         *       *         *  *       -
nw/wsj/07/wsj_0771 0   7       mine   NN   (NP*))))     -    -  5   -         *      *)         *  *    (15)
nw/wsj/07/wsj_0771 0   8          ,    ,         *      -    -  -   -         *       *         *  *       -
nw/wsj/07/wsj_0771 0   9         ''   ''         *)     -    -  -   -         *       *        *)  *       -
nw/wsj/07/wsj_0771 0  10        Mr.  NNP       (NP*     -    -  -   -         *       *     (ARG0*  *     (15
nw/wsj/07/wsj_0771 0  11      Boren  NNP         *)     -    -  -   -  (PERSON)       *         *)  *     15)
nw/wsj/07/wsj_0771 0  12       says  VBZ       (VP*   say   01  1   -         *       *       (V*)  *       -

#end document
```

Data Sample

```
#begin document (nw/wsj/07/wsj_0771); part 000
...
...
nw/wsj/07/wsj_0771 0   0         ``   ``  (TOP(S(S*      -    -  -   -         *       *     (ARG1*  *       -
nw/wsj/07/wsj_0771 0   1 Vandenberg  NNP       (NP*      -    -  -   -  (PERSON)  (ARG1*          *  *  (8|(0)
nw/wsj/07/wsj_0771 0   2        and   CC          *      -    -  -   -         *       *          *  *       -
nw/wsj/07/wsj_0771 0   3    Rayburn  NNP         *)      -    -  -   -  (PERSON)      *)         *  * (23)|8)
nw/wsj/07/wsj_0771 0   4        are  VBP       (VP*     be   01  1   -         *    (V*)          *  *       -
nw/wsj/07/wsj_0771 0   5     heroes  NNS   (NP(NP*)      -    -  -   -         *  (ARG2*          *  *       -
nw/wsj/07/wsj_0771 0   6         of   IN       (PP*      -    -  -   -         *       *          *  *       -
nw/wsj/07/wsj_0771 0   7       mine   NN   (NP*))))      -    -  5   -         *      *)          *  *    (15)
nw/wsj/07/wsj_0771 0   8          ,    ,          *      -    -  -   -         *       *          *  *       -
nw/wsj/07/wsj_0771 0   9         ''   ''         *)      -    -  -   -         *       *         *)  *       -
nw/wsj/07/wsj_0771 0  10        Mr.  NNP       (NP*      -    -  -   -         *       *     (ARG0*  *     (15
nw/wsj/07/wsj_0771 0  11      Boren  NNP         *)      -    -  -   -  (PERSON)       *         *)  *     15)
nw/wsj/07/wsj_0771 0  12       says  VBZ       (VP*    say   01  1   -         *       *       (V*)  *       -

#end document
```

Data Sample

```
#begin document (nw/wsj/07/wsj_0771); part 000
...
...
nw/wsj/07/wsj_0771 0   0         ``   ``  (TOP(S(S*      -    -  -   -         *       *     (ARG1*  *       -
nw/wsj/07/wsj_0771 0   1 Vandenberg  NNP       (NP*      -    -  -   -  (PERSON)  (ARG1*          *  *  (8|(0)
nw/wsj/07/wsj_0771 0   2        and   CC          *      -    -  -   -         *       *          *  *       -
nw/wsj/07/wsj_0771 0   3    Rayburn  NNP         *)      -    -  -   -  (PERSON)      *)         *  * (23)|8)
nw/wsj/07/wsj_0771 0   4        are  VBP       (VP*     be   01  1   -         *    (V*)          *  *       -
nw/wsj/07/wsj_0771 0   5     heroes  NNS   (NP(NP*)      -    -  -   -         *  (ARG2*          *  *       -
nw/wsj/07/wsj_0771 0   6         of   IN       (PP*      -    -  -   -         *       *          *  *       -
nw/wsj/07/wsj_0771 0   7       mine   NN   (NP*))))      -    -  5   -         *      *)          *  *    (15)
nw/wsj/07/wsj_0771 0   8          ,    ,          *      -    -  -   -         *       *          *  *       -
nw/wsj/07/wsj_0771 0   9         ''   ''         *)      -    -  -   -         *       *         *)  *       -
nw/wsj/07/wsj_0771 0  10        Mr.  NNP       (NP*      -    -  -   -         *       *     (ARG0*  *      (15
nw/wsj/07/wsj_0771 0  11      Boren  NNP         *)      -    -  -   -  (PERSON)       *         *)  *      15)
nw/wsj/07/wsj_0771 0  12       says  VBZ       (VP*    say   01  1   -         *       *       (V*)  *       -

#end document
```

Participant Statistics

- **23** participants
- from **11** countries

Country	Participants
Brazil	2
Canada	1
China	6
Germany	3
India	1
Italy	1
Japan	1
Spain	1
Sweden	1
Switzerland	1
USA	5

Official; Closed track; Predicted mentions

System	MD	MUC	B-CUBED	CEAF$_m$	CEAF$_e$	BLANC	Official
	F	F^1	F^2	F	F^3	F	$\frac{F^1+F^2+F^3}{3}$
lee	**70.70**	59.57	68.31	**56.37**	**45.48**	73.02	**57.79**
sapena	43.20	59.55	67.09	53.51	41.32	71.10	55.99
chang	64.28	57.15	**68.79**	54.40	41.94	**73.71**	55.96
nugues	68.96	58.61	65.46	51.45	39.52	71.11	54.53
santos	65.45	56.65	65.66	49.54	37.91	69.46	53.41
song	67.26	**59.95**	63.23	46.29	35.96	61.47	53.05
stoyanov	67.78	58.43	61.44	46.08	35.28	60.28	51.92
sobha	64.23	50.48	64.00	49.48	41.23	63.28	51.90
kobdani	61.03	53.49	65.25	42.70	33.79	62.61	51.04
zhou	62.31	48.96	64.07	47.53	39.74	64.72	50.92
charton	64.30	52.45	62.10	46.22	36.54	64.20	50.36
yang	63.93	52.31	62.32	46.55	35.33	64.63	49.99
hao	64.30	54.47	61.01	45.07	32.67	65.35	49.38
xinxin	61.92	46.62	61.93	44.75	36.23	64.27	48.46
zhang	61.13	47.28	61.14	44.46	35.19	65.21	48.07
kummerfeld	62.72	42.70	60.29	45.35	38.32	59.91	47.10
zhekova	48.29	24.08	61.46	40.43	35.75	53.77	40.43
irwin	26.67	19.98	50.46	31.68	25.21	51.12	31.28

Official; Open track; Predicted mentions

System	MD	MUC	B-CUBED	CEAF_m	CEAF_e	BLANC	Official
	F	F^1	F^2	F	F^3	F	$\frac{F^1+F^2+F^3}{3}$
lee	**70.94**	**61.03**	**68.93**	**56.70**	**44.98**	**73.96**	**58.31**
cai	67.40	57.80	67.66	53.37	41.67	71.62	55.71
uryupina	68.39	57.63	65.18	51.42	40.16	68.88	54.32
klenner	62.28	49.86	64.97	50.03	40.48	69.05	51.77
irwin	35.27	27.21	53.55	33.86	26.76	51.76	35.84

Supplementary; Closed track; Gold boundaries

System	MD	MUC	B-CUBED	CEAF_m	CEAF_e	BLANC	Official
	F	F^1	F^2	F	F^3	F	$\frac{F^1+F^2+F^3}{3}$
lee	**75.16**	**63.90**	**70.03**	**59.26**	**48.30**	**74.77**	**60.74**
nugues	72.42	62.12	66.68	53.84	41.93	71.75	56.91
chang	67.92	59.79	68.65	54.95	41.42	74.29	56.62
santos	67.80	59.52	67.26	51.87	39.72	72.34	55.50
kobdani	66.08	59.57	67.27	44.49	34.92	64.10	53.92
stoyanov	70.29	61.54	62.48	48.08	36.64	62.96	53.55
zhang	64.89	51.64	62.16	46.62	36.95	66.54	50.25
song	66.68	55.48	61.29	43.62	32.53	60.22	49.77
zhekova	62.67	35.22	61.20	41.31	36.38	54.79	44.27

Supplementary; Open track; Gold boundaries

System	MD	MUC	B-CUBED	CEAF_m	CEAF_e	BLANC	Official
	F	F^1	F^2	F	F^3	F	$\frac{F^1+F^2+F^3}{3}$
lee	75.39	65.39	70.78	59.78	47.92	75.83	61.36

Supplementary; Closed track; Gold mentions

System	MD	MUC	B-CUBED	CEAF_m	CEAF_e	BLANC	Official
	F	F^1	F^2	F	F^3	F	$\frac{F^1+F^2+F^3}{3}$
chang	100	82.55	73.70	69.71	65.24	77.26	73.83

Supplementary; Open track; Gold mentions

System	MD	MUC	B-CUBED	CEAF_m	CEAF_e	BLANC	Official
	F	F^1	F^2	F	F^3	F	$\frac{F^1+F^2+F^3}{3}$
lee	90.93	81.56	75.95	70.73	61.64	80.35	73.05

Approaches (I)

	Task	Syn.	Learning Framework	Markable Identification	Markable
lee	C+O	P	Rule-based	Rules to exclude Copular construction, Appositives, Pleonastic *it*, etc.	Feature dependent with shared attributes
sapena	C	P	Decision Tree + Relaxation Labeling	NP (maximal span) + PRP + NE + Capitalized noun heuristic	Full phrase
chang	C	P	Learning Based Java	NP, NE, PRP, PRP$	Full phrase
cai	O	P	Compute hyperedge weights	NP, PRP, PRP$, Base phrase chunks, Pleonastic *it* filter	Full phrase
nugues	C	D	Logistic Regression (LIBLINEAR)	NP, PRP$ and sequence of NNP(s) in post processing using ALIAS and STRINGMATCH	Head word
uryupina	O	P	Decision Tree. Different classifiers for Pro. and non-Pro.	NP, NE, PRP, PRP$, and rules to exclude some specific cases	Full phrase
santos	C	P	ETL committee and Random Forest WEKA)	All NP and all pronouns and PER, ORG, GPE in NP	Full phrase
song	C	P	MaxEnt (OpenNLP)	Mention detection classifier	Full phrase
stoyanov	C	P	Averaged perceptron	NE and possessives in addition to ACE based system	Full phrase
sobha	C	P	CRF for non-pronominal and salience factor for pronouns	Machine learned pleonastic *it*, plus NP, PRP, PRP$ and NE	Minimal (Chunk/NE) and Maximum span
klenner	O	D	Rule-based	NP, NE, PRP, PRP$	Shared attributed/transitivity by using a virtual prototype
kobdani	C	P	Decision Tree	NP (no mention of PRP$)	Start word, End word and Head of NP
zhou	C	P	SVM tree kernel	Rule-based; Five rules: PRP$, PRP, NE, smallest NP subsuming NE and DET+NP	Full phrase
charton	C	P	Multi-layer perceptron	Rules based on POS, NE and filter out pleonastic *it* using rule-based filter	Full phrase
yang	C	P	MaxEnt (MALLET)	NP, PRP, PRP$, pre-modifiers and verbs	Full phrase
hao	C	P	MaxEnt	NP, PRP, PRP$, VBD	Full phrase
xinxin	C	P	ILP/Information gain	NP, PRP, PRP$	Full phrase
zhang	C	P	SVM	IOB classification	Full phrase
kummerfield	C	P	Unsupervised generative model	NP, PRP, PRP$ with maximal span	Full phrase
zhekova	C	P	TIMBL memory based learner	NP, Proper nouns, PRP, PRP$, plus verb with predicate lemma	Head word
irwin	C+O	P	Classification-based ranker	NP, PRP, PRP$	Shared attributes

Approaches (III)

	Verb	Feature Selection	Features	Training
lee	×	×		—
sapena	×	×		Train + Dev
chang	×	×		Train + Dev
cai	×	×		—
nugues	×	Forward + Backward starting from Soon feature set	24	Train + Dev
uryupina	×	Multi-Objective Optimization on three splits. NSGA-II	46	Train + Dev
santos	×	Inherent to the classifiers		Train + Dev
song	×	Same feature set, but per classifier	40	Train
stoyanov	×	×	76	—
sobha	×	×		Train
klenner	×	×		—
kobdani	×	Information gain ratio		Train
zhou	×	×	17	Train + Dev
charton	×	×	22	Train
yang	√	×	40	Train + Dev
hao	√	×		Train + Dev
xinxin	×	Information gain ratio	65	—
zhang	×	×		—
kummerfield	×	×		—
zhekova	√	×		Train + Dev
irwin	×	×		—

Approaches (III)

	Positive Training Examples	Negative Training Examples
lee	—	—
sapena	All mention pairs and longer of nested mentions with common head kept	Mention pairs with less than threshold (5) number of different attribute values are considered (22% out of 99% original are discarded)
chang	Closest antecedent	All preceding mentions in a union of of *gold* and *predicted* mentions. Mentions where the first is pronoun and other not are not considered
cai	Weights are trained on part of the training data	
nugues	Closest Antecedent (Soon, 2001)	Negative examples in between anaphor and closest antecedent (Soon, 2001)
uryupina	Closest antecedent (Soon, 2001)	Negative examples in between anaphor and closest antecedent (Soon, 2001)
santos	Extended version of Soon (2001) where in addition to their strategy, positive and negative examples from mentions in the sentence of the closest preceding antecedent are considered	
song	Pre-cluster pair models separate for each pair NP-NP, NP-PRP and PRP-PRP	
stoyanov	Smart Pair Generation (SmartPG) where the type of antecedent is determined by the type of anaphor using a set of rules	
sobha	Closest antecedent (Soon, 2001)	Negative examples in between anaphor and closest antecedent (Soon, 2001)
klenner	—	—
kobdani	Closest antecedent (Soon, 2001)	Negative examples in between anaphor and closest antecedent (Soon, 2001)
zhou	Closest antecedent (Soon, 2001)	Negative examples in between anaphor and closest antecedent (Soon, 2001)
charton	From the end of the document, until an antecedent is found, or 10 mentions	Negative examples in between anaphor and closest antecedent
yang	Closest antecedent (Soon, 2001)	Negative examples in between anaphor and closest antecedent (Soon, 2001)
hao	Closest antecedent (Soon, 2001)	Negative examples in between anaphor and closest antecedent (Soon, 2001)
xinxin	Closest antecedent (Soon, 2001)	Negative examples in between anaphor and closest antecedent (Soon, 2001)
zhang	Closest antecedent (Soon, 2001)	Negative examples in between anaphor and closest antecedent (Soon, 2001)
kummerfield	—	—
zhekova		Examples in the past three sentences
irwin	Cluster query with NULL cluster for discourse new mentions	

Approaches (IV)

	Decoding	Parse Configuration
lee	Multi-pass Sieves	
sapena	Iterative	1-best
chang	Best link and All links strategy; with and without constraints – Best link without constraints was selected for the official run	
cai	Recursive 2-way Spectral clustering (Agarwal, 2005)	
nugues	Closest-first clustering for pronouns and Best-first clustering for non-pronouns	1-best
uryupina	Mention pair model without ranking as in Soon 2001	
santos	Limited number of preceding mentions 60 for automatic and 40 given gold boundaries; Aggressive-merge clustering (Mccarthy and Lenhert, 1995)	
song	Pre-clusters, with singleton pronoun pre-clusters, and use closest-first clustering. Different link models based on the type of linking mentions – NP-PRP, PRP-PRP and NP-NP	
stoyanov	Single-link clustering by computing transitive closure between pairwise positives.	
sobha	Pronominal: all preceding NPs in the sentence and preceding 4 sentences	
klenner	Incremental entity creation	
kobdani	Best-first clustering. Threshold of 100 words used for long documents	1-best
zhou	—	
charton	MLP with score of 0.5 used for linking and 10 mentions	
yang	Maximum 23 sentences to the left; Constrained clustering	
hao	Beam search (Luo, 2004)	Packed forest
xinxin	Best-first clustering followed by ILP optimization	
zhang	Window of 100 markables	
kummerfield	Rule-based, with Pre- and post- resolution filters	Given + Berkeley parser parses
zhekova	From last possible mention in document	
irwin	Cluster-ranking approach (Rahman and Ng, 2009)	

Conclusions

- Most systems used the **two-pass** mention detection and linking approach
- Very **few participants attempted event** coreference resolution
- System performance seemed to be **stable across genres**
- **Gold standard layer** information did not help much
- **Scoring coreference seems to be a continuing issue**, with very little correlation between various methods
- Choice of **metric makes only small changes to overall ranking**
- Best-performing system (*lee*) was completely **rule-based**

Conclusions

- Most systems used the **two-pass** mention detection and linking approach
- Very **few participants attempted event** coreference resolution
- System performance seemed to be **stable across genres**
- **Gold standard layer** information did not help much
- **Scoring coreference seems to be a continuing issue**, with very little correlation between various methods
- Choice of **metric makes only small changes to overall ranking**
- Best-performing system (*lee*) was completely **rule-based**

Conclusions

- Most systems used the **two-pass** mention detection and linking approach
- Very **few participants attempted event** coreference resolution
- System performance seemed to be **stable across genres**
- **Gold standard layer** information did not help much
- **Scoring coreference seems to be a continuing issue**, with very little correlation between various methods
- Choice of **metric makes only small changes to overall ranking**
- Best-performing system (*lee*) was completely **rule-based**

Conclusions

- Most systems used the **two-pass** mention detection and linking approach
- Very **few participants attempted event** coreference resolution
- System performance seemed to be **stable across genres**
- **Gold standard layer** information did not help much
- **Scoring coreference seems to be a continuing issue**, with very little correlation between various methods
- Choice of **metric makes only small changes to overall ranking**
- Best-performing system (*lee*) was completely **rule-based**

Conclusions

- Most systems used the **two-pass** mention detection and linking approach
- Very **few participants attempted event** coreference resolution
- System performance seemed to be **stable across genres**
- **Gold standard layer** information did not help much
- **Scoring coreference seems to be a continuing issue**, with very little correlation between various methods
- Choice of **metric makes only small changes to overall ranking**
- Best-performing system (*lee*) was completely **rule-based**

Conclusions

- Most systems used the **two-pass** mention detection and linking approach
- Very **few participants attempted event** coreference resolution
- System performance seemed to be **stable across genres**
- **Gold standard layer** information did not help much
- **Scoring coreference seems to be a continuing issue**, with very little correlation between various methods
- Choice of **metric makes only small changes to overall ranking**
- Best-performing system (*lee*) was completely **rule-based**

Conclusions

- Most systems used the **two-pass** mention detection and linking approach
- Very **few participants attempted event** coreference resolution
- System performance seemed to be **stable across genres**
- **Gold standard layer** information did not help much
- **Scoring coreference seems to be a continuing issue**, with very little correlation between various methods
- Choice of **metric makes only small changes to overall ranking**
- Best-performing system (*lee*) was completely **rule-based**

Conclusions

- Most systems used the **two-pass** mention detection and linking approach
- Very **few participants attempted event** coreference resolution
- System performance seemed to be **stable across genres**
- **Gold standard layer** information did not help much
- **Scoring coreference seems to be a continuing issue**, with very little correlation between various methods
- Choice of **metric makes only small changes to overall ranking**
- Best-performing system (*lee*) was completely **rule-based**