Coreference Resolution with World Knowledge

Altaf Rahman and Vincent Ng
Human Language Technology Research Institute
University of Texas at Dallas

Noun Phrase Coreference

- Identify the noun phrases (NPs) that refer to the same real-world entity

Improving Coreference Systems

- Develop new models and methods
- Employ sophisticated linguistic knowledge sources
  - semantic and world knowledge

World Knowledge

- Knowledge about the world that humans use to interpret referring expressions
  - may not be available from the context of a referring expression

Example

Martha Stewart is hoping people don't run out on her.

The celebrity indicted on charges stemming from …

Example

Martha Stewart is hoping people don't run out on her.

The celebrity indicted on charges stemming from …

Example

Martha Stewart is hoping people don't run out on her.

The celebrity indicted on charges stemming from …

- world knowledge has been shown to improve coreference systems

Three Sources of World Knowledge

1. Online encyclopedia and lexical knowledge bases
   - Wikipedia (Ponzetto and Strube, 2006, 2007)
   - YAGO (Bryl et al., 2010; Uryupina et al., 2011)

2. Coreference-annotated data

3. Unannotated data

Goal

- Evaluate commonly-used and under-investigated world knowledge sources for learning-based coreference resolution

Goal

- Evaluate commonly-used and under-investigated world knowledge sources for learning-based coreference resolution

1. Existing work has evaluated a world knowledge source independently of the others
   - do they provide complementary or overlapping knowledge?
   - can they provide further gains when used in combination?

Goal

- Evaluate commonly-used and under-investigated world knowledge sources for learning-based coreference resolution

1. Can they provide further gains when applied in combination?
   - do they offer complementary or overlapping knowledge?

2. Existing work has shown that world knowledge sources can improve the performance of the mention-pair model
   - Can they improve a more sophisticated coreference model
     - e.g., the cluster-ranking model (Rahman & Ng, 2009)?

Goal

- Evaluate commonly-used and under-investigated world knowledge sources for learning-based coreference resolution

1. Can they provide further gains when applied in combination?
   - do they offer complementary or overlapping knowledge?
2. Can they improve a more sophisticated coreference model
   - e.g., the cluster-ranking model (Rahman and Ng, 2009)?
3. Are the gains dependent on the underlying annotation scheme?
   - ACE: coreference among NPs belonging to ACE entity types
   - OntoNotes: “unrestricted” coreference

Three Sources of World Knowledge

1. Online encyclopedia and lexical knowledge bases
   - YAGO
   - FrameNet
2. Coreference-annotated data
3. Unannotated data

Three Sources of World Knowledge

1. Online encyclopedia and lexical knowledge bases
   - YAGO
   - FrameNet
2. Coreference-annotated data
3. Unannotated data

YAGO (Suchanek et al., 2007)

- contains 5 million facts derived from Wikipedia and WordNet
- each fact is a triple describing a relation between two NPs
  - <NP1, rel, NP2>, rel can be one of 90 YAGO relation types

YAGO (Suchanek et al., 2007)

- contains 5 million facts derived from Wikipedia and WordNet
- each fact is a triple describing a relation between two NPs
  - <NP1, rel, NP2>, rel can be one of 90 YAGO relation types
- focuses on two types of YAGO relations: TYPE and MEANS (Bryl et al., 2010, Uryupina et al., 2011)

YAGO (Suchanek et al., 2007)

- contains 5 million facts derived from Wikipedia and WordNet
- each fact is a triple describing a relation between two NPs
  - <NP1, rel, NP2>, rel can be one of 90 YAGO relation types
- focuses on two types of YAGO relations: TYPE and MEANS (Bryl et al., 2010, Uryupina et al., 2011)
  - TYPE: the IS-A relation
    - <AlbertEinstein, TYPE, physicist>
      <BarackObama, TYPE, US president>

YAGO (Suchanek et al., 2007)

- contains 5 million facts derived from Wikipedia and WordNet
- each fact is a triple describing a relation between two NPs
  - <NP1, rel, NP2>, rel can be one of 90 YAGO relation types
- focuses on two types of YAGO relations: TYPE and MEANS (Bryl et al., 2010, Uryupina et al., 2011)
  - TYPE: the IS-A relation
    - <AlbertEinstein, TYPE, physicist>
      <BarackObama, TYPE, US president>
  - MEANS: addresses synonymy and ambiguity
    - <Einstein, MEANS, AlbertEinstein>,
      <Einstein, MEANS, AlfredEinstein>

YAGO (Suchanek et al., 2007)

- contains 5 million facts derived from Wikipedia and WordNet
- each fact is a triple describing a relation between two NPs
  - <NP1, rel, NP2>, rel can be one of 90 YAGO relation types
- focuses on two types of YAGO relations: TYPE and MEANS (Bryl et al., 2010, Uryupina et al., 2011)
  - TYPE: the IS-A relation
    - <AlbertEinstein, TYPE, physicist>
      <BarackObama, TYPE, US president>
  - MEANS: addresses synonymy and ambiguity
    - <Einstein, MEANS, AlbertEinstein>,
      <Einstein, MEANS, AlfredEinstein>
  - provide evidence that the two NPs involved are coreferent

Why YAGO?

- combines the information in Wikipedia and WordNet
- can resolve the celebrity to Martha Stewart
  - neither Wikipedia nor WordNet alone can

Using YAGO for Coreference Resolution

- create a binary-valued YAGO feature
  - Mention-pair model
  - Cluster-ranking model

Using YAGO for Coreference Resolution

- create a binary-valued YAGO feature
  - Mention-pair model
    - determines whether two NPs are coreferent
    - each instance corresponds to two NPs

      { 1 if the two NPs are in a TYPE or MEANS relation
      { 0 otherwise
  - Cluster-ranking model

Using YAGO for Coreference Resolution

- create a binary-valued YAGO feature
  - Mention-pair model
    - determines whether two NPs are coreferent
    - each instance corresponds to two NPs

      1 if the two NPs are in a TYPE or MEANS relation
      0 otherwise
  - Cluster-ranking model
    - ranks coreference clusters preceding each NP to be resolved
    - each instance corresponds to NP_k and a preceding cluster c
    - features are defined between NP_k and c

      1 if NP_k and at least 1 NP in c are in a TYPE or MEANS relation
      0 otherwise

Three Sources of World Knowledge

1. Online encyclopedia and lexical knowledge bases
   - YAGO
   - FrameNet
2. Coreference-annotated data
3. Unannotated data

Motivating Example

Peter Anthony decries program trading as “limiting the game to a few,” but he is not sure whether he wants to denounce it because …

Motivating Example

Peter Anthony decries program trading as "limiting the game to a few," but he is not sure whether he wants to denounce it because …

- To resolve it to program trading, it may be helpful to know
  1. it and program trading have the same semantic role
  2. decry and decounce are "semantically related"

Observation

- Features encoding
  - the semantic roles of the two NPs under consideration
  - whether the associated predicates are “semantically related”

could be useful for identifying coreference relations.

Observation

- Features encoding
  - the semantic roles of the two NPs under consideration
  - whether the associated predicates are "semantically related"

  could be useful for identifying coreference relations.

Use ASSERT

- Provides PropBank-style roles (Arg0, Arg1, …)

Observation

- Features encoding
  - the semantic roles of the two NPs under consideration
  - whether the associated predicates are "semantically related"

  could be useful for identifying coreference relations.

Use ASSERT

- Provides PropBank-style roles (Arg0, Arg1, …)

Use FrameNet

- Checks whether the two predicates appear in the same frame

Observation

- Features encoding
  - the semantic roles of the two NPs under consideration
  - whether the associated predicates are "semantically related"

  could be useful for identifying coreference relations.

Use ASSERT

- Provides PropBank-style roles (Arg0, Arg1, …)

Use FrameNet

- Checks whether the two predicates appear in the same frame
- Consider two verbs related as long as there exists a frame that contains both of them

Features based on FrameNet and ASSERT

Features based on FrameNet and ASSERT

- Assume NP_j and NP_k are the arguments of two predicates

1. Encode knowledge from FrameNet as one of three values
   - The two predicates appear in the same frame
   - Both appear in FrameNet but never in the same frame
   - One or both of them do not appear in FrameNet
2. Encode semantic roles of NP_j and NP_k as one of five values
   - Arg0-Arg0, Arg1-Arg1, Arg0-Arg1, Arg1-Arg0, OTHERS
3. Create 15 binary-valued features by pairing the 3 possible values from FrameNet and 5 possible values from ASSERT

Features based on FrameNet and ASSERT

- Assume NP_j and NP_k are the arguments of two predicates

1. Encode knowledge from FrameNet as one of three values
   - The two predicates appear in the same frame
   - Both appear in FrameNet but never in the same frame
   - One or both of them do not appear in FrameNet
2. Encode semantic roles of NP_j and NP_k as one of five values
   - Arg0-Arg0, Arg1-Arg1, Arg0-Arg1, Arg1-Arg0, OTHERS
3. Create 15 binary-valued features by pairing the 3 possible values from FrameNet and 5 possible values from ASSERT

Features based on FrameNet and ASSERT

- Assume NP_j and NP_k are the arguments of two predicates

1. Encode knowledge from FrameNet as one of three values
   - The two predicates appear in the same frame
   - Both appear in FrameNet but never in the same frame
   - One or both of them do not appear in FrameNet
2. Encode semantic roles of NP_j and NP_k as one of five values
   - Arg0-Arg0, Arg1-Arg1, Arg0-Arg1, Arg1-Arg0, OTHERS
3. Create 15 binary-valued features by pairing the 3 possible values from FrameNet and 5 possible values from ASSERT

Features based on FrameNet and ASSERT

- Assume NP_j and NP_k are the arguments of two predicates

1. Encode knowledge from FrameNet as one of three values
   - The two predicates appear in the same frame
   - Both appear in FrameNet but never in the same frame
   - One or both of them do not appear in FrameNet
2. Encode semantic roles of NP_j and NP_k as one of five values
   - Arg0-Arg0, Arg1-Arg1, Arg0-Arg1, Arg1-Arg0, OTHERS
3. Create 15 binary-valued features by pairing the 3 possible values from FrameNet and 5 possible values from ASSERT

Incorporating Features into Models

- Mention-pair model
  - the 15 features can be employed directly by the mention-pair model, since they are defined on two NPs
- Cluster-ranking model
  - extend their definitions so that they can be computed between an NP and a preceding cluster

Related Work

- No coreference work that employs FrameNet
- But … related to
  - Bean & Riloff's (2004) use of patterns for inducing domain-specific contextual role knowledge
  - Ponzetto & Strube's (2006) use of semantic roles for inducing features

Three Sources of World Knowledge

1. Online encyclopedia and lexical knowledge bases
   - YAGO
   - FrameNet
2. Coreference-annotated data
3. Unannotated data

World Knowledge from Annotated Data

- Observation
  - Since world knowledge is needed for coreference resolution, a human annotator must have employed world knowledge when coreference-annotating a document
- Goal
  - Design features that can “recover” such world knowledge

World Knowledge from Annotated Data

- Observation
  - Since world knowledge is needed for coreference resolution, a human annotator must have employed world knowledge when coreference-annotating a document
- Goal
  - Design features that can “recover” such world knowledge

What kind of world knowledge can we extract from annotated data?

World Knowledge from Annotated Data

1. world knowledge for **identifying coreference relations**
   - if Barack Obama and U.S. president appear in the same coreference chain in a training text, we can gather the world knowledge that Barack Obama is a U.S. president

World Knowledge from Annotated Data

1. world knowledge for **identifying coreference relations**
   - if Barack Obama and U.S. president appear in the same coreference chain in a training text, we can gather the world knowledge that Barack Obama is a U.S. president

2. world knowledge for determining **non-coreference**
   - infer that a lion and a tiger are unlikely to refer to the same entity after realizing that they never appear in the same coreference chain in the training data

World Knowledge from Annotated Data

1. world knowledge for **identifying coreference relations**
   - if Barack Obama and U.S. president appear in the same coreference chain in a training text, we can gather the world knowledge that Barack Obama is a U.S. president

2. world knowledge for determining **non-coreference**
   - infer that a lion and a tiger are unlikely to refer to the same entity after realizing that they never appear in the same coreference chain in the training data
     - features computed based on WordNet distance or distributional similarity may incorrect suggest that the two are coreferent

World Knowledge from Annotated Data

- Observation
  - The NP pairs collected from coreference-annotated training data could be useful features (e.g., <Obama, U.S. president>)

World Knowledge from Annotated Data

- Observation
  - The NP pairs collected from coreference-annotated training data could be useful features (e.g., <Obama, U.S. president>)
- How to compute values for these features?
  - Mention-pair model: feature value is

    $$\begin{cases} 1 & \text{if the feature is composed of the two NPs under consideration} \\ 0 & \text{otherwise} \end{cases}$$

  - Cluster-ranking model
    - Extend this feature definition so that the feature can be applied to an NP and a preceding cluster

World Knowledge from Annotated Data

- Potential problem
  - Data sparsity: many NP pairs in training data may not appear in test data

World Knowledge from Annotated Data

- Potential problem
  - Data sparsity: many NP pairs in training data may not appear in test data
- Solution
  - Employ not only the NP pairs as features but also generalized versions of these features. E.g.,
    - replace a named entity by its named entity tag
    - replace a common NP by its head noun
    - ...

Any Other Useful Knowledge from Annotated Data?

Any Other Useful Knowledge from Annotated Data?

- Recall that … features encoding
  - the semantic roles of two NPs
  - whether the associated verbs are "semantically related"

  could be useful features for coreference resolution

- Goal: create variants of these features

Any Other Useful Knowledge from Annotated Data?

- Recall that … features encoding
  - the semantic roles of two NPs
  - whether the associated verbs are "semantically related"

  could be useful features for coreference resolution

- Goal: create variants of these features

Any Other Useful Knowledge from Annotated Data?

- Recall that … features encoding
  - the semantic roles of two NPs
  - the associated verbs

  could be useful features for coreference resolution

- Goal: create variants of these features

Any Other Useful Knowledge from Annotated Data?

- Recall that … features encoding
  - the semantic roles of two NPs
  - the associated verbs

  could be useful features for coreference resolution

- Goal: create variants of these features

Each feature is represented by two verbs and the semantic roles

- e.g., <decry, denounce, Arg1-Arg1>

Why would these features be useful for coreference?

- They allow a learner to learn from annotated data whether two NPs serving as the objects of *decry* and *denounce* are likely to be coreferent, for instance

Three Sources of World Knowledge

1. Online encyclopedia and lexical knowledge bases
   - YAGO
   - FrameNet

2. Coreference-annotated data

3. Unannotated data

World Knowledge from Unannotated Data

- can extract syntactic appositions heuristically
  - shown to be useful for coreference resolution
    (e.g., Daume & Marcu, 2005, Ng, 2007, Haghighi & Klein, 2009)

- Each extraction is an NP pair. E.g.,
  - <Barack Obama, the president>, …

World Knowledge from Unannotated Data

- can extract syntactic appositions heuristically
  - shown to be useful for coreference resolution
    (e.g., Daume & Marcu, 2005, Ng, 2007, Haghighi & Klein, 2009)
- Each extraction is an NP pair. E.g.,
  - <Barack Obama, the president>, <Delta Airlines, the carrier>
- Create a database consisting of the syntactic appositions extracted from an unannotated corpus
  - 1.057 million NP pairs

Features based on Syntactic Appositions

- Create a binary-valued feature
- Mention-pair model: feature value is
  - 1 if the two NPs appear as a pair in the database
  - 0 otherwise
- Cluster-ranking model
  - extend the definition above so that the feature can be applied to an NP and a preceding cluster

Evaluation

- Evaluate world knowledge sources for coreference resolution

Experimental Setup

- Corpus
  - 410 texts that appear in both OntoNotes-2 and ACE 2004/2005
  - 80% for training, 20% for testing
- NPs extracted automatically
  - ACE: use mention extractor trained on training texts
  - OntoNotes: use Reconcile's markable identification system
- Scoring programs
  - B^3
  - CEAF

Baseline System

- Feature set
  - does not encode world knowledge
  - 39 linguistic features from Rahman & Ng (2009)
- Models
  - trained using linear SVM

B^3 Results on ACE (Baseline)

	Mention-Pair			Cluster-Ranking		
	R	P	F	R	P	F
Baseline	56.5	69.7	**62.4**	61.7	71.2	**66.1**

B^3 Results on ACE (Baseline)

	Mention-Pair			Cluster-Ranking		
	R	P	F	R	P	F
Baseline	56.5	69.7	**62.4**	61.7	71.2	**66.1**

B^3 Results on ACE (Baseline)

	Mention-Pair			Cluster-Ranking		
	R	P	F	R	P	F
Baseline	56.5	69.7	**62.4**	61.7	71.2	**66.1**

B^3 Results on ACE (Baseline)

	Mention-Pair			Cluster-Ranking		
	R	P	F	R	P	F
Baseline	56.5	69.7	**62.4**	61.7	71.2	**66.1**

B^3 Results on ACE (Baseline)

	Mention-Pair			Cluster-Ranking		
	R	P	F	R	P	F
Baseline	56.5	69.7	**62.4**	61.7	71.2	**66.1**

B^3 Results on ACE (Baseline)

	Mention-Pair			Cluster-Ranking		
	R	P	F	R	P	F
Baseline	56.5	69.7	**62.4**	61.7	71.2	**66.1**

- The cluster-ranking model outperforms the mention-pair model

Next …

- Apply the world knowledge sources in isolation to Baseline

Applying World Knowledge Sources in Isolation to Baseline (B^3 Results on ACE)

	Mention-Pair			Cluster-Ranking		
	R	P	F	R	P	F
Baseline	56.5	69.7	**62.4**	61.7	71.2	**66.1**
Baseline+YAGO Type	57.3	70.3	**63.1**	63.5	72.4	**67.6**
Baseline+YAGO Means	56.7	70.0	**62.7**	62.0	71.4	**66.4**
Baseline+Noun Pairs	57.5	70.6	**63.4**	64.1	73.4	**68.4**
Baseline+FrameNet	56.4	70.9	**62.8**	61.8	71.9	**66.5**
Baseline+Verb Pairs	56.9	71.3	**63.3**	62.1	72.2	**66.8**
Baseline+Appositives	56.9	70.0	**62.7**	63.1	71.7	**67.1**

Applying World Knowledge Sources in Isolation to Baseline (B^3 Results on ACE)

	Mention-Pair			Cluster-Ranking		
	R	P	F	R	P	F
Baseline	56.5	69.7	**62.4**	61.7	71.2	**66.1**
Baseline+YAGO Type	57.3	70.3	**63.1**	63.5	72.4	**67.6**
Baseline+YAGO Means	56.7	70.0	**62.7**	62.0	71.4	**66.4**
Baseline+Noun Pairs	57.5	70.6	**63.4**	64.1	73.4	**68.4**
Baseline+FrameNet	56.4	70.9	**62.8**	61.8	71.9	**66.5**
Baseline+Verb Pairs	56.9	71.3	**63.3**	62.1	72.2	**66.8**
Baseline+Appositives	56.9	70.0	**62.7**	63.1	71.7	**67.1**

Applying World Knowledge Sources in Isolation to Baseline (B^3 Results on ACE)

	Mention-Pair			Cluster-Ranking		
	R	P	F	R	P	F
Baseline	56.5	69.7	**62.4**	61.7	71.2	**66.1**
Baseline+YAGO Type	57.3	70.3	**63.1**	63.5	72.4	**67.6**
Baseline+YAGO Means	56.7	70.0	**62.7**	62.0	71.4	**66.4**
Baseline+Noun Pairs	57.5	70.6	**63.4**	64.1	73.4	**68.4**
Baseline+FrameNet	56.4	70.9	**62.8**	61.8	71.9	**66.5**
Baseline+Verb Pairs	56.9	71.3	**63.3**	62.1	72.2	**66.8**
Baseline+Appositives	56.9	70.0	**62.7**	63.1	71.7	**67.1**

Applying World Knowledge Sources in Isolation to Baseline (B^3 Results on ACE)

	Mention-Pair			Cluster-Ranking		
	R	P	F	R	P	F
Baseline	56.5	69.7	**62.4**	61.7	71.2	**66.1**
Baseline+YAGO Type	57.3	70.3	**63.1**	63.5	72.4	**67.6**
Baseline+YAGO Means	56.7	70.0	**62.7**	62.0	71.4	**66.4**
Baseline+Noun Pairs	57.5	70.6	**63.4**	64.1	73.4	**68.4**
Baseline+FrameNet	56.4	70.9	**62.8**	61.8	71.9	**66.5**
Baseline+Verb Pairs	56.9	71.3	**63.3**	62.1	72.2	**66.8**
Baseline+Appositives	56.9	70.0	**62.7**	63.1	71.7	**67.1**

Applying World Knowledge Sources in Isolation to Baseline (B^3 Results on ACE)

	Mention-Pair			Cluster-Ranking		
	R	P	F	R	P	F
Baseline	56.5	69.7	**62.4**	61.7	71.2	**66.1**
Baseline+YAGO Type	57.3	70.3	**63.1**	63.5	72.4	**67.6**
Baseline+YAGO Means	56.7	70.0	**62.7**	62.0	71.4	**66.4**
Baseline+Noun Pairs	57.5	70.6	**63.4**	64.1	73.4	**68.4**
Baseline+FrameNet	56.4	70.9	**62.8**	61.8	71.9	**66.5**
Baseline+Verb Pairs	56.9	71.3	**63.3**	62.1	72.2	**66.8**
Baseline+Appositives	56.9	70.0	**62.7**	63.1	71.7	**67.1**

Applying World Knowledge Sources in Isolation to Baseline (B^3 Results on ACE)

	Mention-Pair			Cluster-Ranking		
	R	P	F	R	P	F
Baseline	56.5	69.7	**62.4**	61.7	71.2	**66.1**
Baseline+YAGO Type	57.3	70.3	**63.1**	63.5	72.4	**67.6**
Baseline+YAGO Means	56.7	70.0	**62.7**	62.0	71.4	**66.4**
Baseline+Noun Pairs	57.5	70.6	**63.4**	64.1	73.4	**68.4**
Baseline+FrameNet	56.4	70.9	**62.8**	61.8	71.9	**66.5**
Baseline+Verb Pairs	56.9	71.3	**63.3**	62.1	72.2	**66.8**
Baseline+Appositives	56.9	70.0	**62.7**	63.1	71.7	**67.1**

Applying World Knowledge Sources in Isolation to Baseline (B^3 Results on ACE)

	Mention-Pair			Cluster-Ranking		
	R	P	F	R	P	F
Baseline	56.5	69.7	**62.4**	61.7	71.2	**66.1**
Baseline+YAGO Type	57.3	70.3	**63.1**	63.5	72.4	**67.6**
Baseline+YAGO Means	56.7	70.0	**62.7**	62.0	71.4	**66.4**
Baseline+Noun Pairs	57.5	70.6	**63.4**	64.1	73.4	**68.4**
Baseline+FrameNet	56.4	70.9	**62.8**	61.8	71.9	**66.5**
Baseline+Verb Pairs	56.9	71.3	**63.3**	62.1	72.2	**66.8**
Baseline+Appositives	56.9	70.0	**62.7**	63.1	71.7	**67.1**

Applying World Knowledge Sources in Isolation to Baseline (B^3 Results on ACE)

	Mention-Pair			Cluster-Ranking		
	R	P	F	R	P	F
Baseline	56.5	69.7	**62.4**	61.7	71.2	**66.1**
Baseline+YAGO Type	57.3	70.3	**63.1**	63.5	72.4	**67.6**
Baseline+YAGO Means	56.7	70.0	**62.7**	62.0	71.4	**66.4**
Baseline+Noun Pairs	57.5	70.6	**63.4**	64.1	73.4	**68.4**
Baseline+FrameNet	56.4	70.9	**62.8**	61.8	71.9	**66.5**
Baseline+Verb Pairs	56.9	71.3	**63.3**	62.1	72.2	**66.8**
Baseline+Appositives	56.9	70.0	**62.7**	63.1	71.7	**67.1**

- Each type of features improves the Baseline for both MP and CR

Applying World Knowledge Sources in Isolation to Baseline (B^3 Results on ACE)

	Mention-Pair			Cluster-Ranking		
	R	P	F	R	P	F
Baseline	56.5	69.7	**62.4**	61.7	71.2	**66.1**
Baseline+YAGO Type	57.3	70.3	**63.1**	63.5	72.4	**67.6**
Baseline+YAGO Means	56.7	70.0	**62.7**	62.0	71.4	**66.4**
Baseline+Noun Pairs	57.5	70.6	**63.4**	64.1	73.4	**68.4**
Baseline+FrameNet	56.4	70.9	**62.8**	61.8	71.9	**66.5**
Baseline+Verb Pairs	56.9	71.3	**63.3**	62.1	72.2	**66.8**
Baseline+Appositives	56.9	70.0	**62.7**	63.1	71.7	**67.1**

- Except for FrameNet, F-score improvements are always accompanied by a simultaneous rise in recall and precision
  - knowledge sources were computed with high accuracies

Applying World Knowledge Sources in Isolation to Baseline (B^3 Results on ACE)

	Mention-Pair			Cluster-Ranking		
	R	P	F	R	P	F
Baseline	56.5	69.7	**62.4**	61.7	71.2	**66.1**
Baseline+YAGO Type	57.3	70.3	**63.1**	63.5	72.4	**67.6**
Baseline+YAGO Means	56.7	70.0	**62.7**	62.0	71.4	**66.4**
Baseline+Noun Pairs	57.5	70.6	**63.4**	64.1	73.4	**68.4**
Baseline+FrameNet	56.4	70.9	**62.8**	61.8	71.9	**66.5**
Baseline+Verb Pairs	56.9	71.3	**63.3**	62.1	72.2	**66.8**
Baseline+Appositives	56.9	70.0	**62.7**	63.1	71.7	**67.1**

- Adding the YAGO Type feature and the Noun Pairs yield the largest improvements over Baseline

Next …

- Add different types of features incrementally to Baseline

Adding Knowledge Sources Incrementally to the Baseline (B^3 Results on ACE)

	Mention-Pair			Cluster-Ranking		
	R	P	F	R	P	F
Baseline	56.5	69.7	**62.4**	61.7	71.2	**66.1**
Baseline+YT	57.3	70.3	**63.1**	63.5	72.4	**67.6**
Baseline+YT+YM	57.8	70.9	**63.6**	63.9	72.6	**68.0**
Baseline+YT+YM+NP	59.5	71.9	**65.1**	66.1	75.4	**70.4**
Baseline+YT+YM+NP+FN	59.6	72.1	**65.3**	66.3	75.1	**70.4**
Baseline+YT+YM+NP+FN+VP	59.9	72.5	**65.6**	66.6	75.9	**70.9**
Baseline+YT+YM+NP+FN+VP+AP	59.7	72.4	**65.4**	66.4	75.7	**70.7**

Adding Knowledge Sources Incrementally to the Baseline (B^3 Results on ACE)

	Mention-Pair			Cluster-Ranking		
	R	P	F	R	P	F
Baseline	56.5	69.7	**62.4**	61.7	71.2	**66.1**
Baseline+YT	57.3	70.3	**63.1**	63.5	72.4	**67.6**
Baseline+YT+YM	57.8	70.9	**63.6**	63.9	72.6	**68.0**
Baseline+YT+YM+NP	59.5	71.9	**65.1**	66.1	75.4	**70.4**
Baseline+YT+YM+NP+FN	59.6	72.1	**65.3**	66.3	75.1	**70.4**
Baseline+YT+YM+NP+FN+VP	59.9	72.5	**65.6**	66.6	75.9	**70.9**
Baseline+YT+YM+NP+FN+VP+AP	59.7	72.4	**65.4**	66.4	75.7	**70.7**

Adding Knowledge Sources Incrementally to the Baseline (B^3 Results on ACE)

	Mention-Pair			Cluster-Ranking		
	R	P	F	R	P	F
Baseline	56.5	69.7	**62.4**	61.7	71.2	**66.1**
Baseline+YT	57.3	70.3	**63.1**	63.5	72.4	**67.6**
Baseline+YT+YM	57.8	70.9	**63.6**	63.9	72.6	**68.0**
Baseline+YT+YM+NP	59.5	71.9	**65.1**	66.1	75.4	**70.4**
Baseline+YT+YM+NP+FN	59.6	72.1	**65.3**	66.3	75.1	**70.4**
Baseline+YT+YM+NP+FN+VP	59.9	72.5	**65.6**	66.6	75.9	**70.9**
Baseline+YT+YM+NP+FN+VP+AP	59.7	72.4	**65.4**	66.4	75.7	**70.7**

Adding Knowledge Sources Incrementally to the Baseline (B^3 Results on ACE)

	Mention-Pair			Cluster-Ranking		
	R	P	F	R	P	F
Baseline	56.5	69.7	**62.4**	61.7	71.2	**66.1**
Baseline+YT	57.3	70.3	**63.1**	63.5	72.4	**67.6**
Baseline+YT+YM	57.8	70.9	**63.6**	63.9	72.6	**68.0**
Baseline+YT+YM+NP	59.5	71.9	**65.1**	66.1	75.4	**70.4**
Baseline+YT+YM+NP+FN	59.6	72.1	**65.3**	66.3	75.1	**70.4**
Baseline+YT+YM+NP+FN+VP	59.9	72.5	**65.6**	66.6	75.9	**70.9**
Baseline+YT+YM+NP+FN+VP+AP	59.7	72.4	**65.4**	66.4	75.7	**70.7**

Adding Knowledge Sources Incrementally to the Baseline (B^3 Results on ACE)

	Mention-Pair			Cluster-Ranking		
	R	P	F	R	P	F
Baseline	56.5	69.7	**62.4**	61.7	71.2	**66.1**
Baseline+YT	57.3	70.3	**63.1**	63.5	72.4	**67.6**
Baseline+YT+YM	57.8	70.9	**63.6**	63.9	72.6	**68.0**
Baseline+YT+YM+NP	59.5	71.9	**65.1**	66.1	75.4	**70.4**
Baseline+YT+YM+NP+FN	59.6	72.1	**65.3**	66.3	75.1	**70.4**
Baseline+YT+YM+NP+FN+VP	59.9	72.5	**65.6**	66.6	75.9	**70.9**
Baseline+YT+YM+NP+FN+VP+AP	59.7	72.4	**65.4**	66.4	75.7	**70.7**

Adding Knowledge Sources Incrementally to the Baseline (B^3 Results on ACE)

	Mention-Pair			Cluster-Ranking		
	R	P	F	R	P	F
Baseline	56.5	69.7	**62.4**	61.7	71.2	**66.1**
Baseline+YT	57.3	70.3	**63.1**	63.5	72.4	**67.6**
Baseline+YT+YM	57.8	70.9	**63.6**	63.9	72.6	**68.0**
Baseline+YT+YM+NP	59.5	71.9	**65.1**	66.1	75.4	**70.4**
Baseline+YT+YM+NP+FN	59.6	72.1	**65.3**	66.3	75.1	**70.4**
Baseline+YT+YM+NP+FN+VP	59.9	72.5	**65.6**	66.6	75.9	**70.9**
Baseline+YT+YM+NP+FN+VP+AP	59.7	72.4	**65.4**	66.4	75.7	**70.7**

Incrementally to the Baseline (B^3 Results on ACE)

	Mention-Pair			Cluster-Ranking		
	R	P	F	R	P	F
Baseline	56.5	69.7	**62.4**	61.7	71.2	**66.1**
Baseline+YT	57.3	70.3	**63.1**	63.5	72.4	**67.6**
Baseline+YT+YM	57.8	70.9	**63.6**	63.9	72.6	**68.0**
Baseline+YT+YM+NP	59.5	71.9	**65.1**	66.1	75.4	**70.4**
Baseline+YT+YM+NP+FN	59.6	72.1	**65.3**	66.3	75.1	**70.4**
Baseline+YT+YM+NP+FN+VP	59.9	72.5	**65.6**	66.6	75.9	**70.9**
Baseline+YT+YM+NP+FN+VP+AP	59.7	72.4	**65.4**	66.4	75.7	**70.7**

- Best result: add all but Appositives to the Baseline

Adding Knowledge Sources Incrementally to the Baseline (B^3 Results on ACE)

	Mention-Pair			Cluster-Ranking		
	R	P	F	R	P	F
Baseline	56.5	69.7	**62.4**	61.7	71.2	**66.1**
Baseline+YT	57.3	70.3	**63.1**	63.5	72.4	**67.6**
Baseline+YT+YM	57.8	70.9	**63.6**	63.9	72.6	**68.0**
Baseline+YT+YM+NP	59.5	71.9	**65.1**	66.1	75.4	**70.4**
Baseline+YT+YM+NP+FN	59.6	72.1	**65.3**	66.3	75.1	**70.4**
Baseline+YT+YM+NP+FN+VP	59.9	72.5	**65.6**	66.6	75.9	**70.9**
Baseline+YT+YM+NP+FN+VP+AP	59.7	72.4	**65.4**	66.4	75.7	**70.7**

- Best result: add all but Appositives to the Baseline
  - F-score increases by 3.2 (MP) and 4.8 (CR) in comparison to Baseline

Adding Knowledge Sources Incrementally to the Baseline (B^3 Results on ACE)

	Mention-Pair			Cluster-Ranking		
	R	P	F	R	P	F
Baseline	56.5	69.7	**62.4**	61.7	71.2	**66.1**
Baseline+YT	57.3	70.3	**63.1**	63.5	72.4	**67.6**
Baseline+YT+YM	57.8	70.9	**63.6**	63.9	72.6	**68.0**
Baseline+YT+YM+NP	59.5	71.9	**65.1**	66.1	75.4	**70.4**
Baseline+YT+YM+NP+FN	59.6	72.1	**65.3**	66.3	75.1	**70.4**
Baseline+YT+YM+NP+FN+VP	59.9	72.5	**65.6**	66.6	75.9	**70.9**
Baseline+YT+YM+NP+FN+VP+AP	59.7	72.4	**65.4**	66.4	75.7	**70.7**

- F-score almost always increases after adding each type of features
  - Different types of features provide complementary knowledge

Next …

- Examine whether the improvements observed in evaluations using ACE annotations carry over to OntoNotes annotations

B^3 Results based on the ACE and OntoNotes Annotation Schemes: Cluster Ranking

	ACE			OntoNotes		
	R	P	F	R	P	F
Baseline	61.7	71.2	**66.1**	59.6	68.8	**63.8**
Baseline+YT	63.5	72.4	**67.6**	61.7	70.0	**65.5**
Baseline+YT+YM	63.9	72.6	**68.0**	62.1	70.4	**66.0**
Baseline+YT+YM+NP	66.1	75.4	**70.4**	62.9	72.4	**67.3**
Baseline+YT+YM+NP+FN	66.3	75.1	**70.4**	63.1	72.3	**67.4**
Baseline+YT+YM+NP+FN+VP	66.6	75.9	**70.9**	63.5	72.9	**67.9**
Baseline+YT+YM+NP+FN+VP+AP	66.4	75.7	**70.7**	63.3	72.9	**67.8**

B^3 Results based on the ACE and OntoNotes Annotation Schemes: Cluster Ranking

	ACE			OntoNotes		
	R	P	F	R	P	F
Baseline	61.7	71.2	**66.1**	59.6	68.8	**63.8**
Baseline+YT	63.5	72.4	**67.6**	61.7	70.0	**65.5**
Baseline+YT+YM	63.9	72.6	**68.0**	62.1	70.4	**66.0**
Baseline+YT+YM+NP	66.1	75.4	**70.4**	62.9	72.4	**67.3**
Baseline+YT+YM+NP+FN	66.3	75.1	**70.4**	63.1	72.3	**67.4**
Baseline+YT+YM+NP+FN+VP	66.6	75.9	**70.9**	63.5	72.9	**67.9**
Baseline+YT+YM+NP+FN+VP+AP	66.4	75.7	**70.7**	63.3	72.9	**67.8**

B^3 Results based on the ACE and OntoNotes Annotation Schemes: Cluster Ranking

	ACE			OntoNotes		
	R	P	F	R	P	F
Baseline	61.7	71.2	**66.1**	59.6	68.8	**63.8**
Baseline+YT	63.5	72.4	**67.6**	61.7	70.0	**65.5**
Baseline+YT+YM	63.9	72.6	**68.0**	62.1	70.4	**66.0**
Baseline+YT+YM+NP	66.1	75.4	**70.4**	62.9	72.4	**67.3**
Baseline+YT+YM+NP+FN	66.3	75.1	**70.4**	63.1	72.3	**67.4**
Baseline+YT+YM+NP+FN+VP	66.6	75.9	**70.9**	63.5	72.9	**67.9**
Baseline+YT+YM+NP+FN+VP+AP	66.4	75.7	**70.7**	63.3	72.9	**67.8**

- Performance trends are similar for both annotation schemes

B^3 Results based on the ACE and OntoNotes Annotation Schemes: Cluster Ranking

	ACE			OntoNotes		
	R	P	F	R	P	F
Baseline	61.7	71.2	**66.1**	59.6	68.8	**63.8**
Baseline+YT	63.5	72.4	**67.6**	61.7	70.0	**65.5**
Baseline+YT+YM	63.9	72.6	**68.0**	62.1	70.4	**66.0**
Baseline+YT+YM+NP	66.1	75.4	**70.4**	62.9	72.4	**67.3**
Baseline+YT+YM+NP+FN	66.3	75.1	**70.4**	63.1	72.3	**67.4**
Baseline+YT+YM+NP+FN+VP	66.6	75.9	**70.9**	63.5	72.9	**67.9**
Baseline+YT+YM+NP+FN+VP+AP	66.4	75.7	**70.7**	63.3	72.9	**67.8**

- Performance trends are similar for both annotation schemes
  - Best results achieved by adding all but Appositives to Baseline

B^3 Results based on the ACE and OntoNotes Annotation Schemes: Cluster Ranking

	ACE			OntoNotes		
	R	P	F	R	P	F
Baseline	61.7	71.2	**66.1**	59.6	68.8	**63.8**
Baseline+YT	63.5	72.4	**67.6**	61.7	70.0	**65.5**
Baseline+YT+YM	63.9	72.6	**68.0**	62.1	70.4	**66.0**
Baseline+YT+YM+NP	66.1	75.4	**70.4**	62.9	72.4	**67.3**
Baseline+YT+YM+NP+FN	66.3	75.1	**70.4**	63.1	72.3	**67.4**
Baseline+YT+YM+NP+FN+VP	66.6	75.9	**70.9**	63.5	72.9	**67.9**
Baseline+YT+YM+NP+FN+VP+AP	66.4	75.7	**70.7**	63.3	72.9	**67.8**

- Performance trends are similar for both annotation schemes
  - Best results achieved by adding all but Appositives to Baseline
  - F-score almost always increases after adding each feature type

Summary

- Evaluated different sources of world knowledge when used by the mention-pair model and the cluster-ranking model
  - each type of features improves Baseline when used in isolation
  - all but the Appositive features improve F-score when added incrementally to the Baseline
  - performance trends remain the same regardless of the underlying coreference model and annotation scheme
  - while each type of features provides small gains, their cumulative benefits are substantial