Translation-Based Projection for Multilingual Coreference Resolution

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

Noun Phrase Coreference

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

- Lots of work on English coreference, but there has also been work on coreference in other languages

Multilingual Coreference Resolution

- A natural next step
- An important next step
  - Coreference resolvers do not exist for many languages

Multilingual Coreference Resolution

- A natural next step
- An important next step
  - Coreference resolvers do not exist for many languages
- Surge of interest in multilingual coreference resolution
  - ACE 2004/2005
    - English, Chinese, Arabic
  - SemEval 2010 Task 1
    - English, Spanish, Catalan, Italian, Dutch, German
  - CoNLL 2012 shared task
    - English, Chinese, Arabic

Multilingual Coreference Resolution: How?

- We have coreference-annotated data for multiple languages
  - Employ a supervised approach
    - Train a coreference resolver for each language

Multilingual Coreference Resolution: How?

- We have coreference-annotated data for multiple languages
  - Employ a supervised approach
    - Train a coreference resolver for each language

- **Weakness**: corpus annotation bottleneck
  - For each new language of interest, need to coreference-annotate a potentially large number of documents

How about a Rule-Based Approach?

- Revived interest in rule-based approaches owing to the Stanford resolver’s competitive performance

- **Strength**: no need to coreference-annotate any data

How about a Rule-Based Approach?

- Revived interest in rule-based approaches owing to the Stanford resolver’s competitive performance

- **Strength**: no need to coreference-annotate any data

- **Weakness**: we are replacing the corpus annotation bottleneck with the knowledge acquisition bottleneck
  - Need knowledge of the target language to design rules

How about an Unsupervised Approach?

- Markov logic networks? (Poon & Domingoes, 2008)
  - Need to write coreference rules

How about an Unsupervised Approach?

- Markov logic networks? (Poon & Domingoes, 2008)
  - Need to write coreference rules

- Generative models? (Haghighi & Klein, 2010; Ng, 2008)
  - Need linguistic knowledge to design the generative story and combine the knowledge sources

How about an Unsupervised Approach?

- Markov logic networks? (Poon & Domingoes, 2008)
  - Need to write coreference rules

- Generative models? (Haghighi & Klein, 2010; Ng, 2008)
  - Need linguistic knowledge to design the generative story and combine the knowledge sources

- Unsupervised coreference models are not models that can be designed without knowledge of the target language

But … we still need to pick an approach

- Argument for a heuristic/unsupervised approach:
  - Designing coreference rules and generative models may not be as time-consuming as coreference-annotating data

But … we still need to pick an approach

- Argument for a heuristic/unsupervised approach:
  - Designing coreference rules and generative models may not be as time-consuming as coreference-annotating data

- This may be true for English
  - can easily write a rule to enforce gender/number agreement

But … we still need to pick an approach

- Argument for a heuristic/unsupervised approach:
    - Designing coreference rules and generative models may not be as time-consuming as coreference-annotating data
- This may be true for English
    - can easily write a rule to enforce gender/number agreement
- But .. computing these features may not be simple for …
    - Chinese
        - No morphology
            - difficult to determine number
        - Many first names used by both gender
            - difficult to determine gender

Annotated Data are indispensible

- But given the high cost of coreference-annotating data, need to obtain annotated data in a cost-effective manner

Annotated Data are indispensible

- But given the high cost of coreference-annotating data, need to obtain annotated data in a cost-effective manner

Translation-based projection

Translation-Based Projection

Source language

Target language

Translation-Based Projection

Source language

- coreference resolver available

Target language

Translation-Based Projection

Source language

- coreference resolver available

Target language

- coreference resolver not available

Translation-Based Projection

Source language

- coreference resolver available

Target language

- coreference resolver not available

- Goal
  - coreference-annotate documents in target language using resolver in source language

Translation-Based Projection

Source language	**Target language**
coreference resolver available	coreference resolver not available

- Goal
  - coreference-annotate documents in target language using resolver in source language
- Idea
  - project annotations produced by resolver from source to target

Translation-Based Projection Algorithm

- Input: document in target language
- Output: document coreference-annotated

Translation-Based Projection Algorithm

- Input: document in target language
- Output: document coreference-annotated

- 3 steps:
  1. Machine-translate document from target to source
  2. Run resolver on the translated document
  3. Project annotations from source back to target

Translation-Based Projection: Example

玛丽告诉约翰她非常喜欢他。

Translation-Based Projection: Example

玛丽告诉约翰她非常喜欢他。

1. Machine-translate document from target to source

Translation-Based Projection: Example

玛丽告诉约翰她非常喜欢他。

Mary told John that she liked him a lot.

1. Machine-translate document from target to source

Translation-Based Projection: Example

玛丽告诉约翰她非常喜欢他。

Mary told John that she liked him a lot.

2. Run resolver on the translated document
   - to extract mentions and produce coreference chains

Translation-Based Projection: Example

玛丽告诉约翰她非常喜欢他。

Mary told John that she liked him a lot.

2. Run resolver on the translated document
   - to extract mentions and produce coreference chains

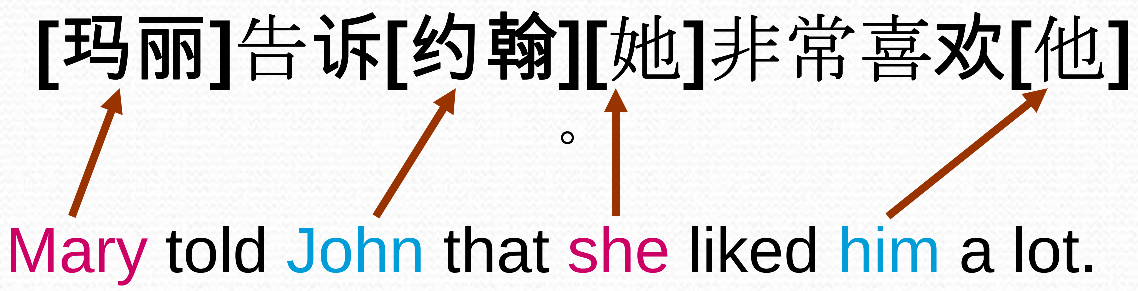
Translation-Based Projection: Example

玛丽告诉约翰她非常喜欢他。

Mary told John that she liked him a lot.

3. Project annotations from source back to target
   - project mentions

Translation-Based Projection: Example

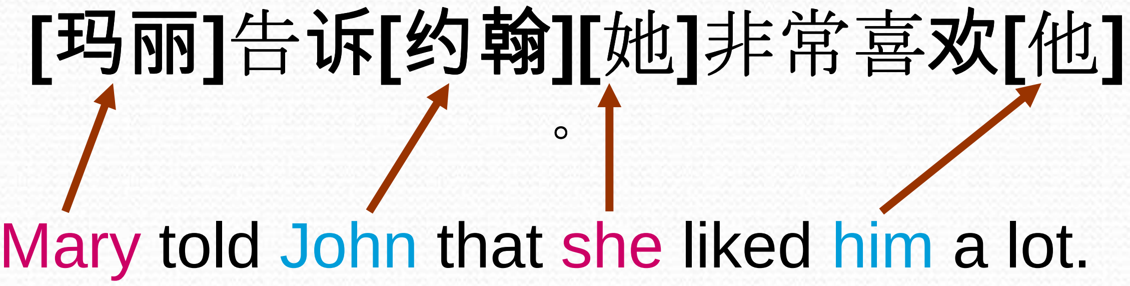


[玛丽]告诉[约翰][她]非常喜欢[他]
。

Mary told John that she liked him a lot.

3. Project annotations from source back to target
   - project mentions

Translation-Based Projection: Example

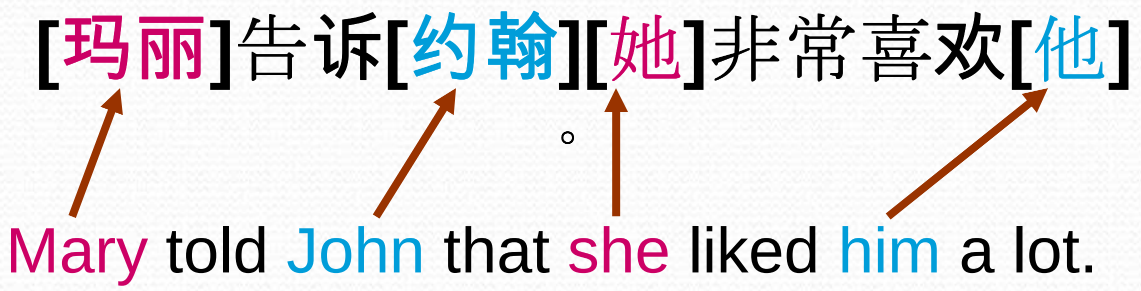


[玛丽]告诉[约翰][她]非常喜欢[他]。

Mary told John that she liked him a lot.

3. Project annotations from source back to target
   - project mentions
   - project coreference chains

Translation-Based Projection: Example



[玛丽]告诉[约翰][她]非常喜欢[他]。

Mary told John that she liked him a lot.

3. Project annotations from source back to target
   - project mentions
   - project coreference chains

Translation-Based Projection

- No corpus annotation bottleneck
- No knowledge acquisition bottleneck

- Problem solved?

Translation-Based Projection

- No corpus annotation bottleneck
- No knowledge acquisition bottleneck

- Problem solved?
  - Not really

Translation-Based Projection

- No corpus annotation bottleneck
- No knowledge acquisition bottleneck

- Problem solved?
  - Not really
  - Projection is not a solution to multilingual coreference problem
    - Every language has its own idiosyncrasies
    - Projection cannot produce annotations capturing language-specific properties
      - E.g., zero pronouns

Goal

- Explore the extent to which projection can push the limits of multilingual coreference resolution
  - projection is not meant to and cannot replace corpus annotation

Caveat

- Translation-based projection won’t work if MT service for the target language is not available

Caveat

- Translation-based projection won’t work if MT service for the target language is not available

- True, but …
  - Number of language pairs for which MT services are available is increasing
  - Parallel corpus may be used, if available

Plan for the Talk

- Translation-based projection
  - Related work
  - Implementation details
  - Evaluation

Plan for the Talk

- Translation-based projection
  - Related work
  - Implementation details
  - Evaluation

Related Work

- Projecting annotations from a resource-rich language to a resource-poor language
  - proposed by Yarowsky and Ngai (2001)
  - assumes a parallel corpus for the source and target languages
  - more recent work uses an MT engine instead

Related Work

- Applying projection to coreference resolution
  - Idea formulated in a declined EU proposal circa 2005
  - Postolache et al. (2006)
    - English-Romanian parallel corpus: Orwell's "1984"
    - Manually create coreference annotations on English side
    - Automatically project English annotations to Romanian
    - Manually fix projection errors
  - Harabagiu and Maiorano (2000)
    - English-Romanian parallel corpus: manually translating MUC-6
    - Manually project MUC-6 coreference annotations to Romanian

Related Work

- So … their goal is different from ours
  - They create clean coreference corpus by employing significant knowledge of the target language
  - We create a coreference corpus via an entirely automatic process without using knowledge of the target language

Plan for the Talk

- Translation-based projection
  - Related work
  - Implementation details
  - Evaluation

Translation-Based Projection

Source language **Target language**

Translation-Based Projection

(resource-rich)

Source language

(resource-poor)

Target language

Translation-Based Projection

(resource-rich)

Source language

- coreference resolver available

(resource-poor)

Target language

- coreference resolver not available

Translation-Based Projection

(resource-rich)

Source language

- coreference resolver available

(resource-poor)

Target language

- coreference resolver not available
- not necessarily resource-poor
  - we may have many linguistic taggers at the morphological, syntactic and semantic levels

Translation-Based Projection

(resource-rich)

Source language

- coreference resolver available

(resource-poor)

Target language

- coreference resolver not available
- not necessarily resource-poor
  - we may have many linguistic taggers at the morphological, syntactic and semantic levels

- Goal
  - Examine whether the linguistic taggers for the target language, if available, can be exploited to improve projection approach

Translation-Based Projection

- Evaluate the projection approach in 3 settings
  - Differ in terms of the extent to which linguistic taggers for the target language are available

Translation-Based Projection

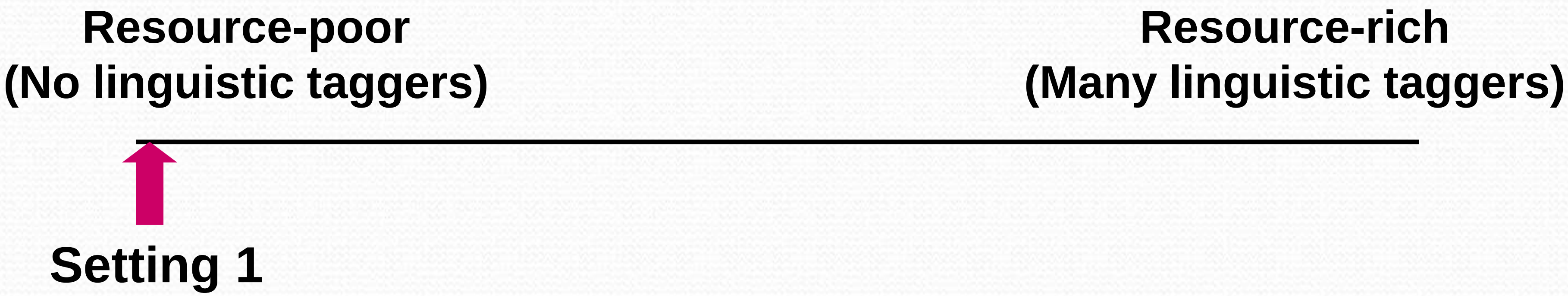
- Evaluate the projection approach in 3 settings
  - Differ in terms of the extent to which linguistic taggers for the target language are available

Resource-poor
(No linguistic taggers)

Resource-rich
(Many linguistic taggers)

Translation-Based Projection

- Evaluate the projection approach in 3 settings
  - Differ in terms of the extent to which linguistic taggers for the target language are available



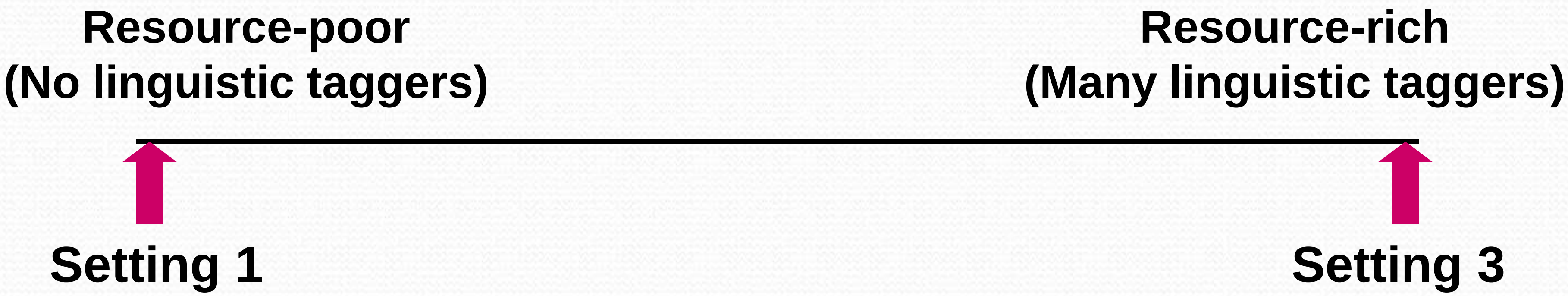
Resource-poor
(No linguistic taggers)

Resource-rich
(Many linguistic taggers)

Setting 1

Translation-Based Projection

- Evaluate the projection approach in 3 settings
  - Differ in terms of the extent to which linguistic taggers for the target language are available



Resource-poor
(No linguistic taggers)

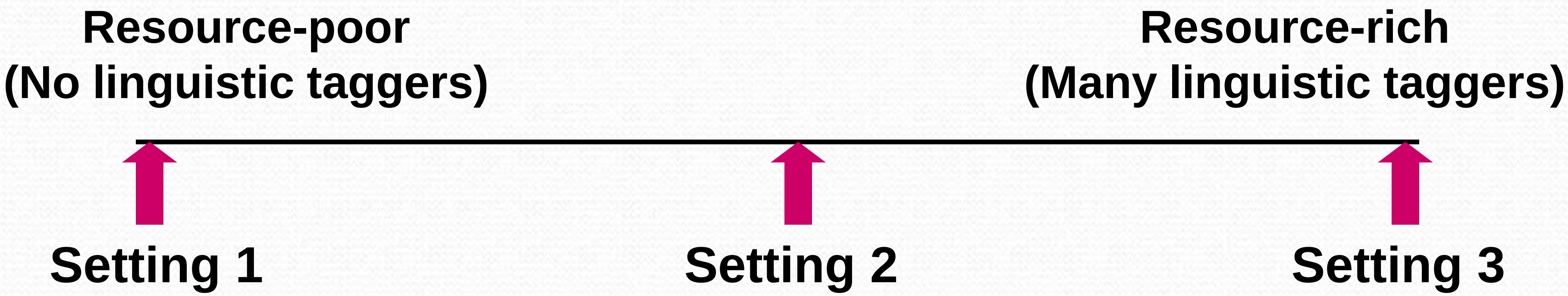
Resource-rich
(Many linguistic taggers)

Setting 1

Setting 3

Translation-Based Projection

- Evaluate the projection approach in 3 settings
  - Differ in terms of the extent to which linguistic taggers for the target language are available



Resource-poor
(No linguistic taggers)

Resource-rich
(Many linguistic taggers)

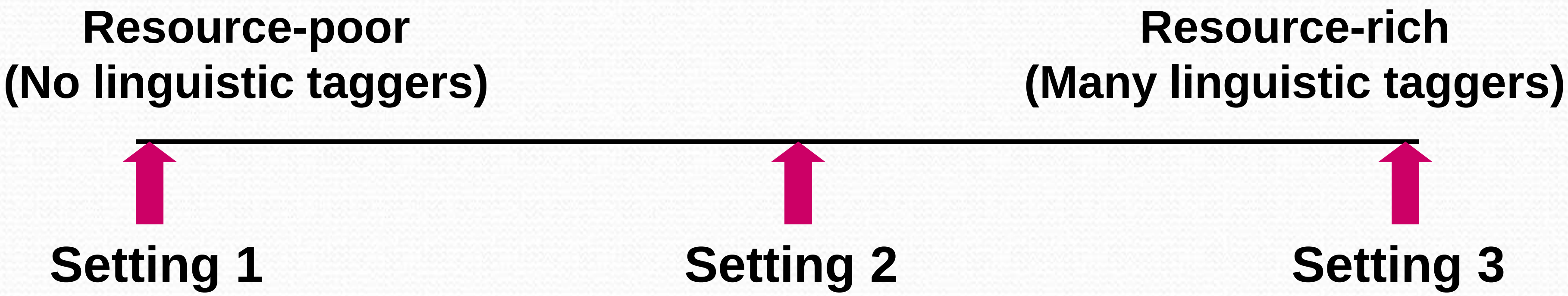
Setting 1

Setting 2

Setting 3

Translation-Based Projection

- Evaluate the projection approach in 3 settings
  - Differ in terms of the extent to which linguistic taggers for the target language are available



Resource-poor
(No linguistic taggers)

Resource-rich
(Many linguistic taggers)

Setting 1

Setting 2

Setting 3

  - Assume **English** is source language
    Chinese is target language

Setting 1

- No Chinese taggers available
- Simply apply the 3 steps involved in MT-based projection

1. Machine-translate text from Chinese to English

2. Run resolver on the translated English text

3. Project annotations from English text back to Chinese

Setting 1

- No Chinese taggers available
- Simply apply the 3 steps involved in MT-based projection
  1. Machine-translate text from Chinese to English
     - GoogleTranslate
  2. Run resolver on the translated English text
  3. Project annotations from English text back to Chinese

Setting 1

- No Chinese taggers available
- Simply apply the 3 steps involved in MT-based projection
  1. Machine-translate text from Chinese to English
     - GoogleTranslate
  2. Run resolver on the translated English text
     - Reconcile (mention detection and coreference resolution)
  3. Project annotations from English text back to Chinese

Setting 1

- No Chinese taggers available
- Simply apply the 3 steps involved in MT-based projection
  1. Machine-translate text from Chinese to English
     - GoogleTranslate
  2. Run resolver on the translated English text
     - Reconcile (mention detection and coreference resolution)
  3. Project annotations from English text back to Chinese
     - GIZA++ for Chinese-to-English word alignment
     - Heuristically create Chinese mentions from Reconcile mentions

Setting 1

- No Chinese taggers available
- Simply apply the 3 steps involved in MT-based projection
  1. Machine-translate text from Chinese to English
     - GoogleTranslate
  2. Run resolver on the translated English text
     - Reconcile (mention detection and coreference resolution)
  3. Project annotations from English text back to Chinese
     - GIZA++ for Chinese-to-English word alignment
       - Improve alignment via a bilingual dictionary from web sources
     - Heuristically create Chinese mentions from Reconcile mentions

Setting 1

- No Chinese taggers available
- Simply apply the 3 steps involved in MT-based projection
  1. Machine-translate text from Chinese to English
     - GoogleTranslate
  2. Run resolver on the translated English text
     - Reconcile (mention detection and coreference resolution)
  3. Project annotations from English text back to Chinese
     - GIZA++ for Chinese-to-English word alignment
       - Improve alignment via a bilingual dictionary from web sources
     - Heuristically create Chinese mentions from Reconcile mentions
       - Use Yarowsky and Ngai's (2001) NP projection method

Setting 2

- A Chinese mention detector available

- How can we profitably exploit this mention detector?

Setting 2

玛丽告诉约翰她非常喜欢他。

Setting 2

玛丽告诉约翰她非常喜欢他。

1. Apply Chinese mention detector to extract mentions

[玛丽]告诉[约翰][她]非常喜欢[他] 。

Setting 2

玛丽告诉约翰她非常喜欢他。

1. Apply Chinese mention detector to extract mentions

   [玛丽]告诉[约翰][她]非常喜欢[他] 。

2. Machine-translate text to English (using GoogleTranslate)

   Mary told John that she liked him a lot.

Setting 2

玛丽告诉约翰她非常喜欢他。

1. Apply Chinese mention detector to extract mentions

   [玛丽]告**诉[约翰][**她**]**非常喜**欢[**他**]** 。

2. Machine-translate text to English (using GoogleTranslate)

   Mary told John that she liked him a lot.

3. Use Reconcile to detect mentions and perform coreference

   [Mary] told [John] that [she] liked [him] a lot.

Setting 2

玛丽告诉约翰她非常喜欢他。

1. Apply Chinese mention detector to extract mentions

   [玛丽]告**诉[约翰][**她**]**非常喜**欢[**他**]** 。

2. Machine-translate text to English (using GoogleTranslate)

   Mary told John that she liked him a lot.

3. Use Reconcile to detect mentions and perform coreference

   [Mary] told [John] that [she] liked [him] a lot.

4. Project chains back to Chinese (using word alignment)

   [玛丽]告**诉[约翰][**她**]**非常喜**欢[**他**]** 。

Setting 2

- What’s the difference between Setting 2 and Setting 1?

Setting 2

- What’s the difference between Setting 2 and Setting 1?

Setting 2	
• Chinese mentions detected using Chinese mention detector • English mentions detected using Reconcile	

Setting 2

- What's the difference between Setting 2 and Setting 1?

Setting 2	Setting 1
• Chinese mentions detected using Chinese mention detector • English mentions detected using Reconcile	• Chinese mentions projected from English mentions • Chinese mention boundaries defined by NP projection algorithm (English mentions and word alignment)

Setting 2

- What’s the difference between Setting 2 and Setting 1?

Setting 2	Setting 1
• Chinese mentions detected using Chinese mention detector • English mentions detected using Reconcile	• Chinese mentions projected from English mentions • Chinese mention boundaries defined by NP projection algorithm (English mentions and word alignment)

Mention boundaries on neither side depends on word alignment

Setting 2

- What's the difference between Setting 2 and Setting 1?

Setting 2	Setting 1
• Chinese mentions detected using Chinese mention detector • English mentions detected using Reconcile	• Chinese mentions projected from English mentions • Chinese mention boundaries defined by NP projection algorithm (English mentions and word alignment)
Mention boundaries on neither side depends on word alignment	Chinese mention boundaries are sensitive to word alignment errors.

Setting 2

- What's the difference between Setting 2 and Setting 1?

Setting 2	Setting 1
• Chinese mentions detected using Chinese mention detector • English mentions detected using Reconcile	• Chinese mentions projected from English mentions • Chinese mention boundaries defined by NP projection algorithm (English mentions and word alignment)

Mention boundaries on neither side depends on word alignment

Chinese mention boundaries are sensitive to word alignment errors.

Setting 2's mention detection method more robust to word alignment errors

Setting 3

- Additional linguistic taggers for Chinese (e.g., NE taggers, semantic taggers) available

- How can we profitably exploit these Chinese taggers?

Setting 3

- Additional linguistic taggers for Chinese (e.g., NE taggers, semantic taggers) available

- How can we profitably exploit these Chinese taggers?
  - Use them to generate features to train a Chinese coreference resolver in a supervised manner

Setting 3

- Additional linguistic taggers for Chinese (e.g., NE taggers, semantic taggers) available

- How can we profitably exploit these Chinese taggers?
  - Use them to generate features to train a Chinese coreference resolver in a supervised manner
    - But we don’t have any manual coreference annotations to train a supervised resolver

Setting 3

- Additional linguistic taggers for Chinese (e.g., NE taggers, semantic taggers) available

- How can we profitably exploit these Chinese taggers?
  - Use them to generate features to train a Chinese coreference resolver in a supervised manner
    - But we don’t have any manual coreference annotations to train a supervised resolver
    - Idea (Kobdani et al., 2011):
      - use pseudo coreference annotations
      - Setting 2 can be used to produce these pseudo annotations

Setting 3

- In our experiments, we didn’t run any linguistic taggers for the target language
  - Took a shared task dataset for the target language
    - Features have been computed for each word in the dataset
    - Partition the dataset into a training set and a test set
    - Train a coreference resolver on training set by replacing correct coreference labels with pseudo labels generated via Setting 2

Setting 3

- In our experiments, we didn’t run any linguistic taggers for the target language
    - Took a shared task dataset for the target language
        - Features have been computed for each word in the dataset
        - Partition the dataset into a training set and a test set
        - Train a coreference resolver on training set by replacing correct coreference labels with pseudo labels generated via Setting 2

- Setting 3
    - exploits information provided by additional taggers
        - but no manual coreference annotations are needed

Plan for the Talk

- Translation-based projection
  - Related work
  - Implementation details
  - Evaluation

Data Sets

- Italian and Spanish datasets from SemEval-2010 shared task on Coreference Resolution in Multiple Languages
  - Each dataset is composed of a training set and a test set
  - Statistics:

	Italian		Spanish	
	Training	**Test**	**Training**	**Test**
Number of mentions	24853	13394	78779	14133
Number of non-singleton clusters	18376	9520	48681	8789
Number of singleton clusters	15984	8288	37336	6737

Scoring Programs

- 4 scoring programs used in the shared task
  - MUC (Vilain et al., 1995)
  - B^3 (Bagga and Baldwin, 1998)
  - ϕ_3-CEAF (Luo, 1995)
  - BLANC (Recasens and Hovy, 2011)

Gold vs. Regular Settings

- refer to what information in a dataset can be used
- format of dataset follows that of a CoNLL shared task dataset
  - each row corresponds to a word
  - each column corresponds to a feature
    - some correspond to manually computed features
    - some correspond to automatically computed features

Gold vs. Regular Settings

- refer to what information in a dataset can be used
- format of dataset follows that of a CoNLL shared task dataset
  - each row corresponds to a word
  - each column corresponds to a feature
    - some correspond to manually computed features
    - some correspond to automatically computed features

Gold setting **Regular setting**

Gold vs. Regular Settings

- refer to what information in a dataset can be used
- format of dataset follows that of a CoNLL shared task dataset
  - each row corresponds to a word
  - each column corresponds to a feature
    - some correspond to manually computed features
    - some correspond to automatically computed features

Gold setting

- Use gold mentions and manually computed features

Regular setting

- Use automatically computed mentions and features

Results for Italian

	Regular (CEAF)			Gold (CEAF)		
	R	P	F	R	P	F
Supervised	73.7	74.3	**74.0**	74.5	74.5	**74.5**

Results for Italian

	Regular (CEAF)			Gold (CEAF)		
	R	P	F	R	P	F
Supervised	73.7	74.3	**74.0**	74.5	74.5	**74.5**

Results for Italian

	Regular (CEAF)			Gold (CEAF)		
	R	P	F	R	P	F
Supervised	73.7	74.3	**74.0**	74.5	74.5	**74.5**

Results for Italian

	Regular (CEAF)			Gold (CEAF)		
	R	P	F	R	P	F
Supervised	73.7	74.3	**74.0**	74.5	74.5	**74.5**

Results for Italian

	Regular (CEAF)			Gold (CEAF)		
	R	P	F	R	P	F
Supervised	73.7	74.3	**74.0**	74.5	74.5	**74.5**

Results for Italian

	Regular (CEAF)			Gold (CEAF)		
	R	P	F	R	P	F
Supervised	73.7	74.3	**74.0**	74.5	74.5	**74.5**

- obtained via a resolver trained on all training data using all features made available by the shared task organizers
- upper bound on performance of our projection approach

Results for Italian

	Regular (CEAF)			Gold (CEAF)		
	R	P	F	R	P	F
Supervised	73.7	74.3	**74.0**	74.5	74.5	**74.5**
Best result in Shared Task	57.1	66.2	**61.3**	66.0	66.0	**66.0**

Results for Italian

	Regular (CEAF)			Gold (CEAF)		
	R	P	F	R	P	F
Supervised	73.7	74.3	**74.0**	74.5	74.5	**74.5**
Best result in Shared Task	57.1	66.2	**61.3**	66.0	66.0	**66.0**

- Sanity check on whether upper bounds established by our supervised resolver are reasonable

Results for Italian

	Regular (CEAF)			Gold (CEAF)		
	R	P	F	R	P	F
Supervised	73.7	74.3	**74.0**	74.5	74.5	**74.5**
Best result in Shared Task	57.1	66.2	**61.3**	66.0	66.0	**66.0**

- Results worse than those of our supervised resolver

Results for Italian

	Regular (CEAF)			Gold (CEAF)		
	R	P	F	R	P	F
Supervised	73.7	74.3	**74.0**	74.5	74.5	**74.5**
Best result in Shared Task	57.1	66.2	**61.3**	66.0	66.0	**66.0**
1. No linguistic taggers	17.0	26.0	**20.6**	---	---	**---**

Results for Italian

	Regular (CEAF)			Gold (CEAF)		
	R	P	F	R	P	F
Supervised	73.7	74.3	**74.0**	74.5	74.5	**74.5**
Best result in Shared Task	57.1	66.2	**61.3**	66.0	66.0	**66.0**
1. No linguistic taggers	17.0	26.0	**20.6**	---	---	**---**

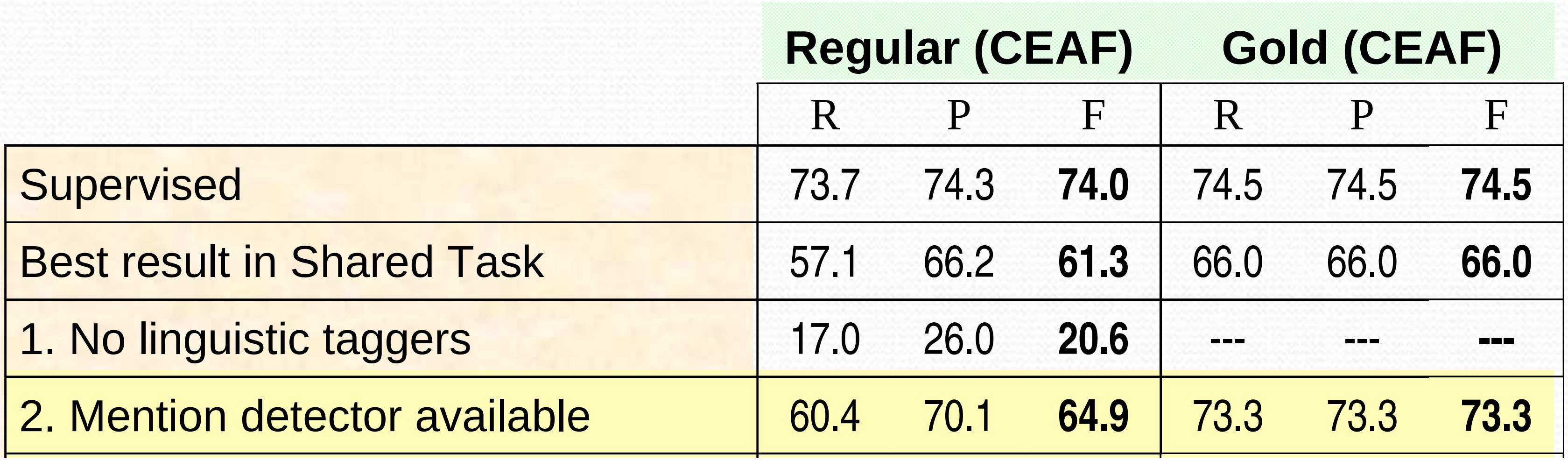
- No gold results
  - No gold mentions or manually computed features are used
  - Mentions are projected from the Reconcile mentions

Results for Italian

	Regular (CEAF)			Gold (CEAF)		
	R	P	F	R	P	F
Supervised	73.7	74.3	**74.0**	74.5	74.5	**74.5**
Best result in Shared Task	57.1	66.2	**61.3**	66.0	66.0	**66.0**
1. No linguistic taggers	17.0	26.0	**20.6**	---	---	**---**

- F-score significantly worse than its supervised counterparts ($p < 0.05$, paired t-test)

Results for Italian



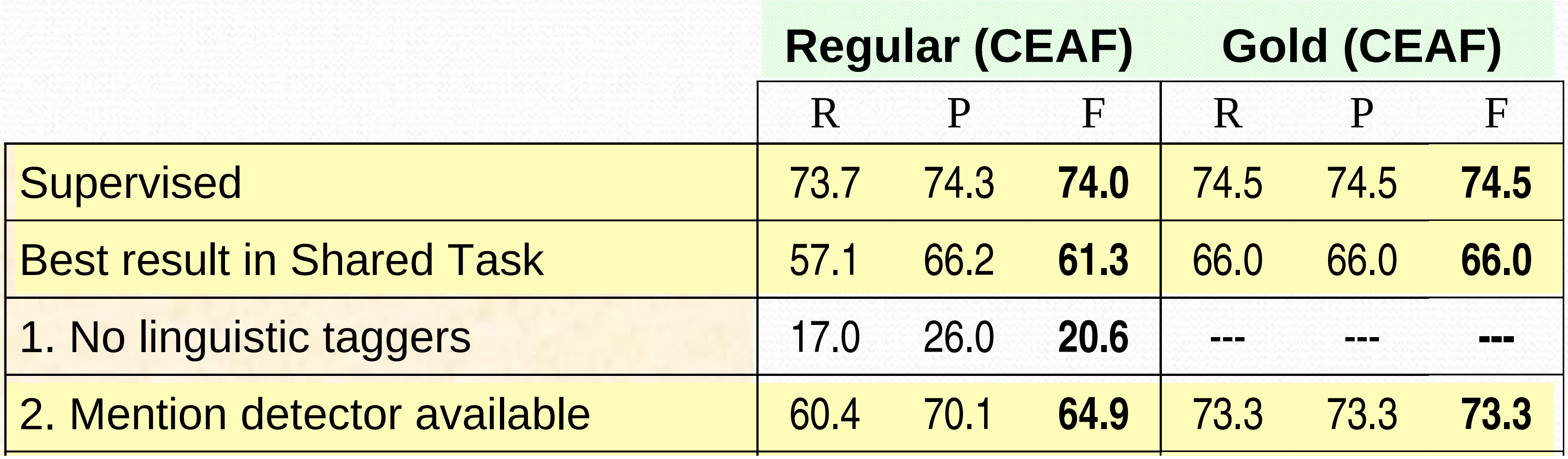
	Regular (CEAF)			Gold (CEAF)		
	R	P	F	R	P	F
Supervised	73.7	74.3	**74.0**	74.5	74.5	**74.5**
Best result in Shared Task	57.1	66.2	**61.3**	66.0	66.0	**66.0**
1. No linguistic taggers	17.0	26.0	**20.6**	---	---	**---**
2. Mention detector available	60.4	70.1	**64.9**	73.3	73.3	**73.3**

Results for Italian

	Regular (CEAF)			Gold (CEAF)		
	R	P	F	R	P	F
Supervised	73.7	74.3	**74.0**	74.5	74.5	**74.5**
Best result in Shared Task	57.1	66.2	**61.3**	66.0	66.0	**66.0**
1. No linguistic taggers	17.0	26.0	**20.6**	---	---	**---**
2. Mention detector available	60.4	70.1	**64.9**	73.3	73.3	**73.3**

- In comparison to Setting 1
  - Setting 2 yields significantly better results

Results for Italian



	Regular (CEAF)			Gold (CEAF)		
	R	P	F	R	P	F
Supervised	73.7	74.3	**74.0**	74.5	74.5	**74.5**
Best result in Shared Task	57.1	66.2	**61.3**	66.0	66.0	**66.0**
1. No linguistic taggers	17.0	26.0	**20.6**	---	---	**---**
2. Mention detector available	60.4	70.1	**64.9**	73.3	73.3	**73.3**

- In comparison to supervised results:
  - beats best shared task resolver
  - lags behind our supervised resolver

Results for Italian

	Regular (CEAF)			Gold (CEAF)		
	R	P	F	R	P	F
Supervised	73.7	74.3	**74.0**	74.5	74.5	**74.5**
Best result in Shared Task	57.1	66.2	**61.3**	66.0	66.0	**66.0**
1. No linguistic taggers	17.0	26.0	**20.6**	---	---	**---**
2. Mention detector available	60.4	70.1	**64.9**	73.3	73.3	**73.3**
3. Additional taggers available	61.1	62.9	**61.9**	64.3	64.3	**64.3**

Results for Italian

	Regular (CEAF)			**Gold (CEAF)**		
	R	P	F	R	P	F
Supervised	73.7	74.3	**74.0**	74.5	74.5	**74.5**
Best result in Shared Task	57.1	66.2	**61.3**	66.0	66.0	**66.0**
1. No linguistic taggers	17.0	26.0	**20.6**	---	---	**---**
2. Mention detector available	60.4	70.1	**64.9**	73.3	73.3	**73.3**
3. Additional taggers available	61.1	62.9	**61.9**	64.3	64.3	**64.3**

- In comparison to Setting 2
  - Performance drops for both Regular and Gold settings

Italian Results

	Regular (CEAF)			Regular (MUC)		
	R	P	F	R	P	F
Supervised	73.7	74.3	**74.0**	31.9	68.0	**43.4**
Best result in Shared Task	57.1	66.2	**61.3**	50.1	50.7	**50.4**
1. No linguistic taggers	17.0	26.0	**20.6**	8.1	28.5	**12.6**
2. Mention detector available	60.4	70.1	**64.9**	17.2	68.2	**27.5**
3. Additional taggers available	61.1	62.9	**61.9**	29.5	63.2	**40.2**

Italian Results

	Regular (CEAF)			Regular (MUC)		
	R	P	F	R	P	F
Supervised	73.7	74.3	**74.0**	31.9	68.0	**43.4**
Best result in Shared Task	57.1	66.2	**61.3**	50.1	50.7	**50.4**
1. No linguistic taggers	17.0	26.0	**20.6**	8.1	28.5	**12.6**
2. Mention detector available	60.4	70.1	**64.9**	17.2	68.2	**27.5**
3. Additional taggers available	61.1	62.9	**61.9**	29.5	63.2	**40.2**

Italian Results

	Regular (CEAF)			Regular (MUC)		
	R	P	F	R	P	F
Supervised	73.7	74.3	**74.0**	31.9	68.0	**43.4**
Best result in Shared Task	57.1	66.2	**61.3**	50.1	50.7	**50.4**
1. No linguistic taggers	17.0	26.0	**20.6**	8.1	28.5	**12.6**
2. Mention detector available	60.4	70.1	**64.9**	17.2	68.2	**27.5**
3. Additional taggers available	61.1	62.9	**61.9**	29.5	63.2	**40.2**

Italian Results

	Regular (CEAF)			Regular (MUC)		
	R	P	F	R	P	F
Supervised	73.7	74.3	**74.0**	31.9	68.0	**43.4**
Best result in Shared Task	57.1	66.2	**61.3**	50.1	50.7	**50.4**
1. No linguistic taggers	17.0	26.0	**20.6**	8.1	28.5	**12.6**
2. Mention detector available	60.4	70.1	**64.9**	17.2	68.2	**27.5**
3. Additional taggers available	61.1	62.9	**61.9**	29.5	63.2	**40.2**

- CEAF: Setting 3 is worse than Setting 2 (poorer precision)
- MUC: Setting 3 is better than Setting 2 (better recall)

Italian Results

	Regular (CEAF)			Regular (MUC)		
	R	P	F	R	P	F
Supervised	73.7	74.3	**74.0**	31.9	68.0	**43.4**
Best result in Shared Task	57.1	66.2	**61.3**	50.1	50.7	**50.4**
1. No linguistic taggers	17.0	26.0	**20.6**	8.1	28.5	**12.6**
2. Mention detector available	60.4	70.1	**64.9**	17.2	68.2	**27.5**
3. Additional taggers available	61.1	62.9	**61.9**	29.5	63.2	**40.2**

- Setting 3 has higher recall according to both scoring programs
  - More coreference links are discovered
  - The additional taggers have enabled us to discover new links

Italian Results

	Regular (CEAF)			Regular (MUC)		
	R	P	F	R	P	F
Supervised	73.7	74.3	**74.0**	31.9	68.0	**43.4**
Best result in Shared Task	57.1	66.2	**61.3**	50.1	50.7	**50.4**
1. No linguistic taggers	17.0	26.0	**20.6**	8.1	28.5	**12.6**
2. Mention detector available	60.4	70.1	**64.9**	17.2	68.2	**27.5**
3. Additional taggers available	61.1	62.9	**61.9**	29.5	63.2	**40.2**

- Setting 3 has higher recall according to both scoring programs
  - More coreference links are discovered
  - The additional taggers have enabled us to discover new links
- Setting 3 has lower precision according to both scoring programs
  - Some of these new links are spurious

Italian Results

	Regular (CEAF)			Regular (MUC)		
	R	P	F	R	P	F
Supervised	73.7	74.3	**74.0**	31.9	68.0	**43.4**
Best result in Shared Task	57.1	66.2	**61.3**	50.1	50.7	**50.4**
1. No linguistic taggers	17.0	26.0	**20.6**	8.1	28.5	**12.6**
2. Mention detector available	60.4	70.1	**64.9**	17.2	68.2	**27.5**
3. Additional taggers available	61.1	62.9	**61.9**	29.5	63.2	**40.2**

- MUC gives a much higher recall to Setting 3 than to Setting 2
- CEAF gives only a slightly higher recall to Setting 3

Italian Results

	Regular (CEAF)			Regular (MUC)		
	R	P	F	R	P	F
Supervised	73.7	74.3	**74.0**	31.9	68.0	**43.4**
Best result in Shared Task	57.1	66.2	**61.3**	50.1	50.7	**50.4**
1. No linguistic taggers	17.0	26.0	**20.6**	8.1	28.5	**12.6**
2. Mention detector available	60.4	70.1	**64.9**	17.2	68.2	**27.5**
3. Additional taggers available	61.1	62.9	**61.9**	29.5	63.2	**40.2**

- MUC gives a much higher recall to Setting 3 than to Setting 2
- CEAF gives only a slightly higher recall to Setting 3
  - MUC scores only coreference links, not singleton clusters
  - CEAF scores both coreference links and singleton clusters

Italian Results

	Regular (CEAF)			Regular (MUC)		
	R	P	F	R	P	F
Supervised	73.7	74.3	**74.0**	31.9	68.0	**43.4**
Best result in Shared Task	57.1	66.2	**61.3**	50.1	50.7	**50.4**
1. No linguistic taggers	17.0	26.0	**20.6**	8.1	28.5	**12.6**
2. Mention detector available	60.4	70.1	**64.9**	17.2	68.2	**27.5**
3. Additional taggers available	61.1	62.9	**61.9**	29.5	63.2	**40.2**

- MUC gives a much higher recall to Setting 3 than to Setting 2
- CEAF gives only a slightly higher recall to Setting 3
  - MUC scores only coreference links, not singleton clusters
  - CEAF scores both coreference links and singleton clusters
    - overwhelmed by the large number singleton clusters

Results (Averaged over all Scoring Programs)

	Italian		Spanish	
	Regular	Gold	Regular	Gold
	F	F	F	F
Supervised	63.4	65.9	54.6	66.1
Best result in Shared Task	60.0	61.2	49.6	66.8
Setting 1	21.4	---	37.6	---
Setting 2	54.9	58.2	46.8	56.1
Setting 3	57.7	58.9	51.7	61.4

Results (Averaged over all Scoring Programs)

	Italian		Spanish	
	Regular	**Gold**	**Regular**	**Gold**
	F	F	F	F
Supervised	63.4	65.9	54.6	66.1
Best result in Shared Task	60.0	61.2	49.6	66.8
Setting 1	21.4	---	37.6	---
Setting 2	54.9	58.2	46.8	56.1
Setting 3	57.7	58.9	51.7	61.4

Results (Averaged over all Scoring Programs)

	Italian		Spanish	
	Regular	Gold	Regular	Gold
	F	F	F	F
Supervised	63.4	65.9	54.6	66.1
Best result in Shared Task	60.0	61.2	49.6	66.8
Setting 1	21.4	---	37.6	---
Setting 2	54.9	58.2	46.8	56.1
Setting 3	57.7	58.9	51.7	61.4

Results (Averaged over all Scoring Programs)

	Italian		Spanish	
	Regular	Gold	Regular	Gold
	F	F	F	F
Supervised	63.4	65.9	54.6	66.1
Best result in Shared Task	60.0	61.2	49.6	66.8
Setting 1	21.4	---	37.6	---
Setting 2	54.9	58.2	46.8	56.1
Setting 3	57.7	58.9	51.7	61.4

- Supervised results comparable to/better than shared task result

Results (Averaged over all Scoring Programs)

	Italian		Spanish	
	Regular	Gold	Regular	Gold
	F	F	F	F
Supervised	63.4	65.9	54.6	66.1
Best result in Shared Task	60.0	61.2	49.6	66.8
Setting 1	21.4	---	37.6	---
Setting 2	54.9	58.2	46.8	56.1
Setting 3	57.7	58.9	51.7	61.4

- Setting 2 results are better than Setting 1 results

Results (Averaged over all Scoring Programs)

	Italian		Spanish	
	Regular	Gold	Regular	Gold
	F	F	F	F
Supervised	63.4	65.9	54.6	66.1
Best result in Shared Task	60.0	61.2	49.6	66.8
Setting 1	21.4	---	37.6	---
Setting 2	54.9	58.2	46.8	56.1
Setting 3	57.7	58.9	51.7	61.4

- Setting 3 results are
  - slightly better than Setting 2 results for Italian
  - significantly better than Setting 2 results for Spanish

Results (Averaged over all Scoring Programs)

	Italian		Spanish	
	Regular	Gold	Regular	Gold
	F	F	F	F
Supervised	63.4	65.9	54.6	66.1
Best result in Shared Task	60.0	61.2	49.6	66.8
Setting 1	21.4	---	37.6	---
Setting 2	54.9	58.2	46.8	56.1
Setting 3	57.7	58.9	51.7	61.4

- Setting 3 results are around 89-94% of the supervised results

Results (Averaged over all Scoring Programs)

	Italian		Spanish	
	Regular	Gold	Regular	Gold
	F	F	F	F
Supervised	63.4	65.9	54.6	66.1
Best result in Shared Task	60.0	61.2	49.6	66.8
Setting 1	21.4	---	37.6	---
Setting 2	54.9	58.2	46.8	56.1
Setting 3	57.7	58.9	51.7	61.4

- Setting 3 results are around 89-94% of the supervised results
  - obtained without any manual coreference annotations

Summary

- Investigated MT-based projection approach to coreference
  - can perform coreference resolution for a language
    - without coreference-annotated data
    - without linguistic knowledge of the language
  - can exploit any available knowledge about the target language

- Obtained promising results for Italian and Spanish
  - achieved ~90% of the performance of a supervised resolver when only a mention detector for the target language is available

- Has the potential to allow coreference technologies to be deployed across a larger number of languages

Future Work

- Isolate the impact of each factor that harms performance
  - Errors in MT, coreference in source language, projection

- Explore alternatives
  - Translate all coreference-annotated data from source to target, then train a coreference model on the translated data

- Use our approach to alleviate corpus annotation bottleneck
  - Use the annotated data it produces to augment the manual coreference annotations capturing language-specific properties