Cross-Lingual Cross-Document Coreference with Entity Linking

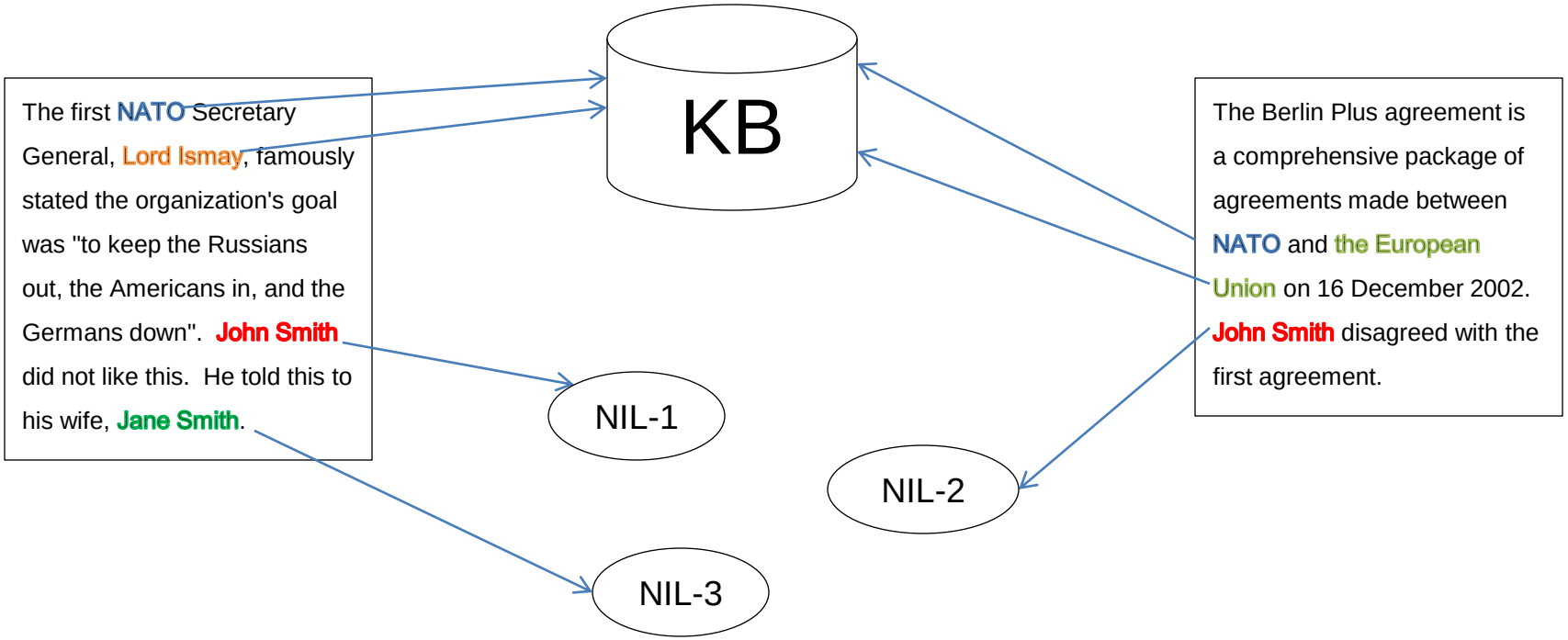
Sean Monahan, John Lehmann,

Timothy Nyberg, Jesse Plymale, Arnold Jung

NIST

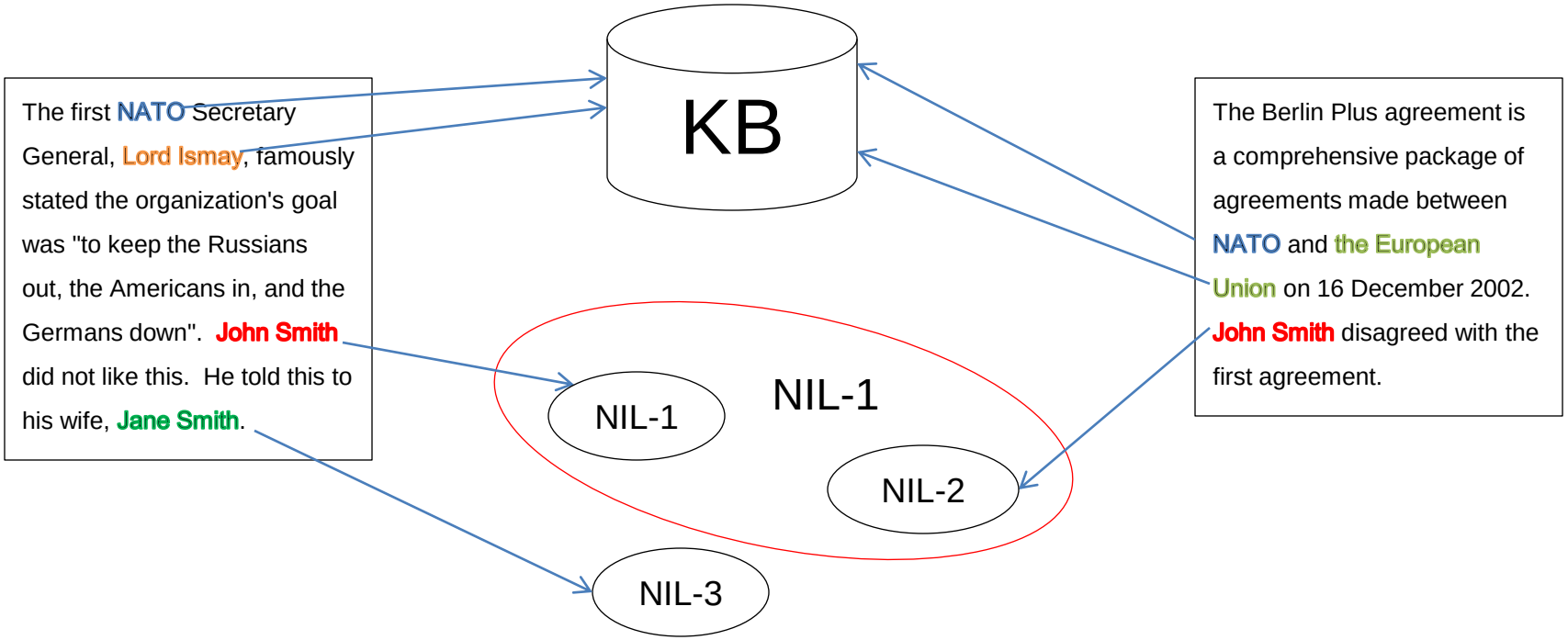
2010 Entity Linking Task

- Link entity mentions in text to Knowledge Base (KB)
  - Each entity mention is given a KB identifier
  - *Non-clustering linker*



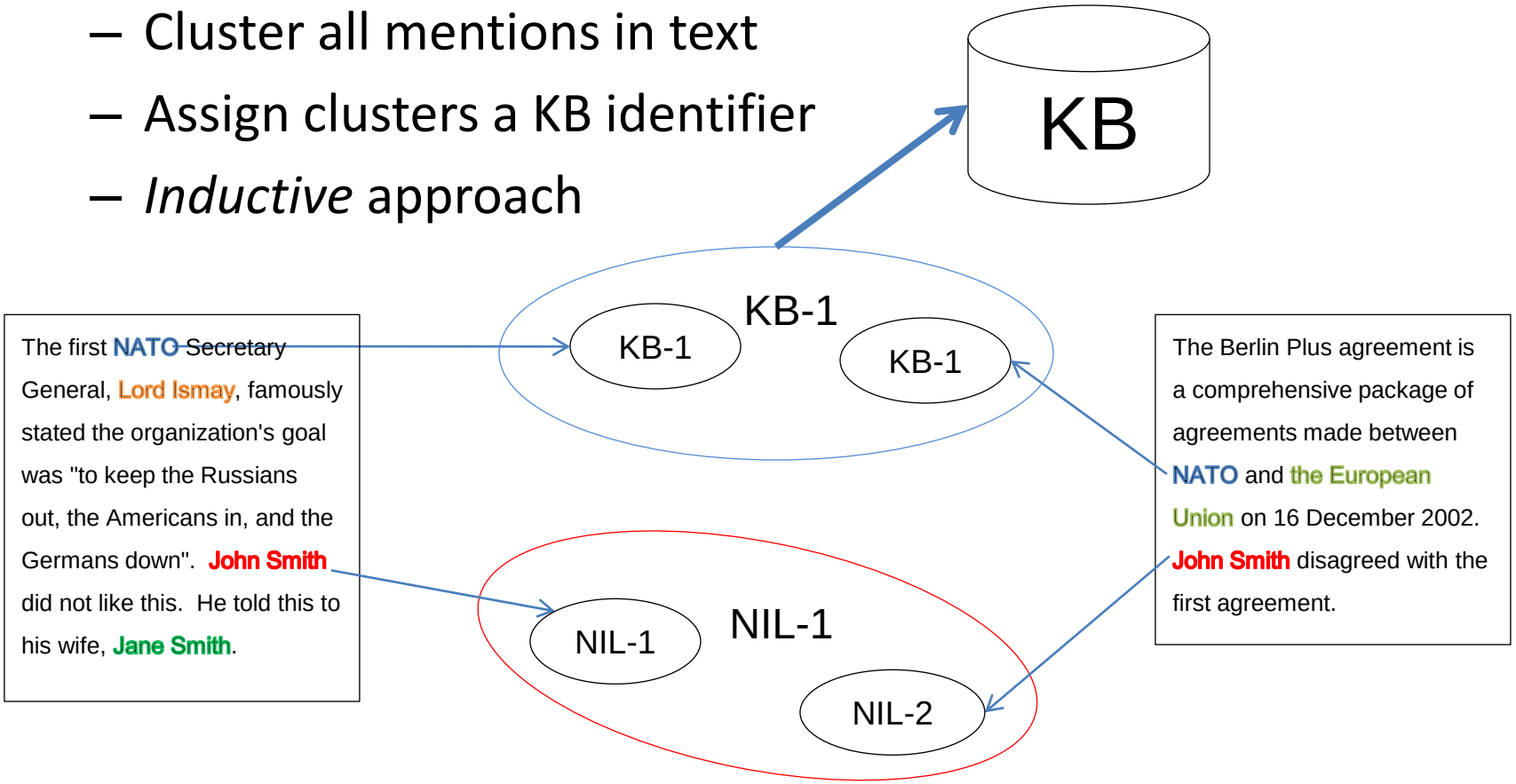
2011 Entity Linking with NIL Clustering Task

- Additionally, cluster all of the remaining NILs
  - Perhaps the most important entities might be the ones you haven't heard of yet
- *Deductive* approach: First link, then cluster remaining NILs



2011 Entity Linking with NIL Clustering Task

- Alternate view: Cross-Document Coreference (CDC) approach
  - Cluster all mentions in text
  - Assign clusters a KB identifier
  - *Inductive* approach



Talk Overview

1. English Entity Linking (with NIL Clustering)
   - Made extensive use of 2010 Entity Linking System
     - Details in (Lehmann et al., 2010)
   - Focus on extending task to NIL clustering
     - 4-stage clustering algorithm
     - Show that our method:
       - Successfully performs NIL clustering
       - Improves linking accuracy on non-NIL entities
   - Improvements to 2010 entity linking algorithm (non-clustering)

Talk Overview (cont.)

2. Cross-Lingual Entity Linking with NIL Clustering
   - Two Approaches
     - Native Language Entity Linking
     - Translation with English Linking

2011 Entity Linking with NIL Clustering Components

LANGUAGE COMPUTER

- Necessary components
  1. **Synonymy**
     - Determine entities likely to match
     - “National Security Council” → “NSC”
  2. **Polysemy**
     - Extract features and cluster similar entities
     - “NSC” (Iran) ≠ “NSC” (Malaysia)
  3. **KB Linking / NIL Detection**
     - Decide between the best KB identifier and NIL for each cluster

NIST

Approach

0. Preprocess each document
   - Includes **entity links** using the *non-clustering linker*
1. Group by similar names
2. Resolve polysemy with agglomerative clustering
3. Resolve synonymy by merging clusters
4. Link each cluster to the knowledge base

CDC: Stage 1

Group by similar names

- Has effect of splitting languages

1

"We and other countries have expressed our concern to the Chinese," said a spokesman for the National Security Council, Gordon Johndroe.

Iran's National Security Council has announced that it will "suspend" the releasing of 15 British sailors and marines detained by Iranian forces on March 23.

The document "reflects the broad interagency effort under way in Iraq" according to an NSC spokesman Frederick Jones

CDC: Stage 2

Cluster within the groups to resolve polysemy

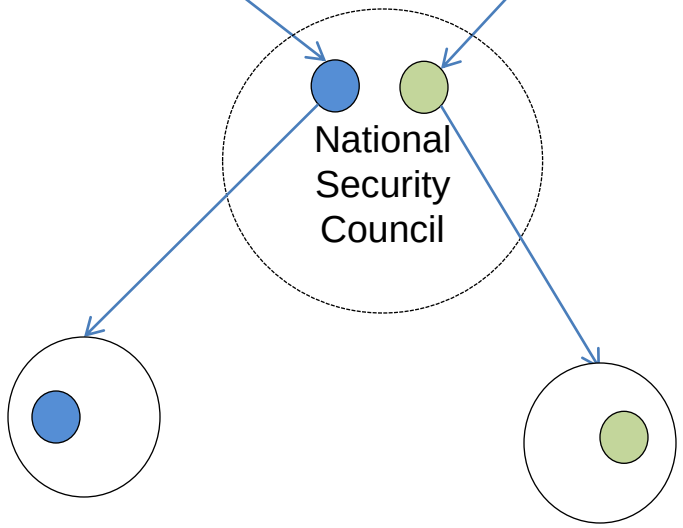
"We and other countries have expressed our concern to the Chinese," said a spokesman for the National Security Council, Gordon Johndroe.

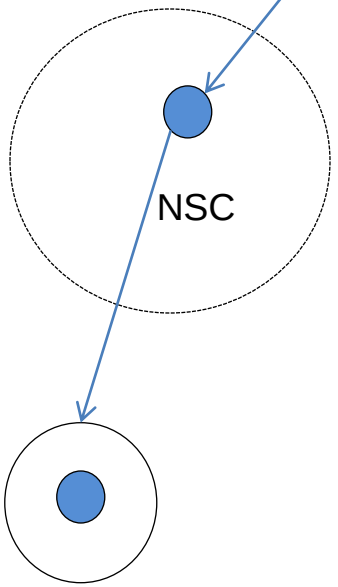
Iran's National Security Council has announced that it will "suspend" the releasing of 15 British sailors and marines detained by Iranian forces on March 23.

The document "reflects the broad interagency effort under way in Iraq" according to an NSC spokesman Frederick Jones

1

2

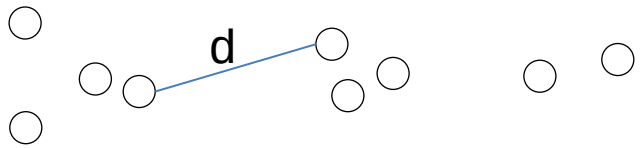


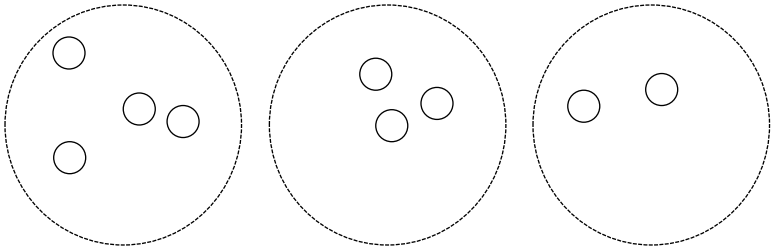


CDC: Stage 2 Clustering Algorithm

Supervised hierarchical agglomerative clustering

- (Gooi and Allan, 1998)
- Balanced Data Set (Akbani et al., 2004)





$$d(M_1, M_2) = \frac{1}{|M_1| \cdot |M_2|} \sum_{m_1 \in M_1} \sum_{m_2 \in M_2} d(m_1, m_2)$$

$$merge\ if\ d < \tau$$

CDC: Stage 2 Features

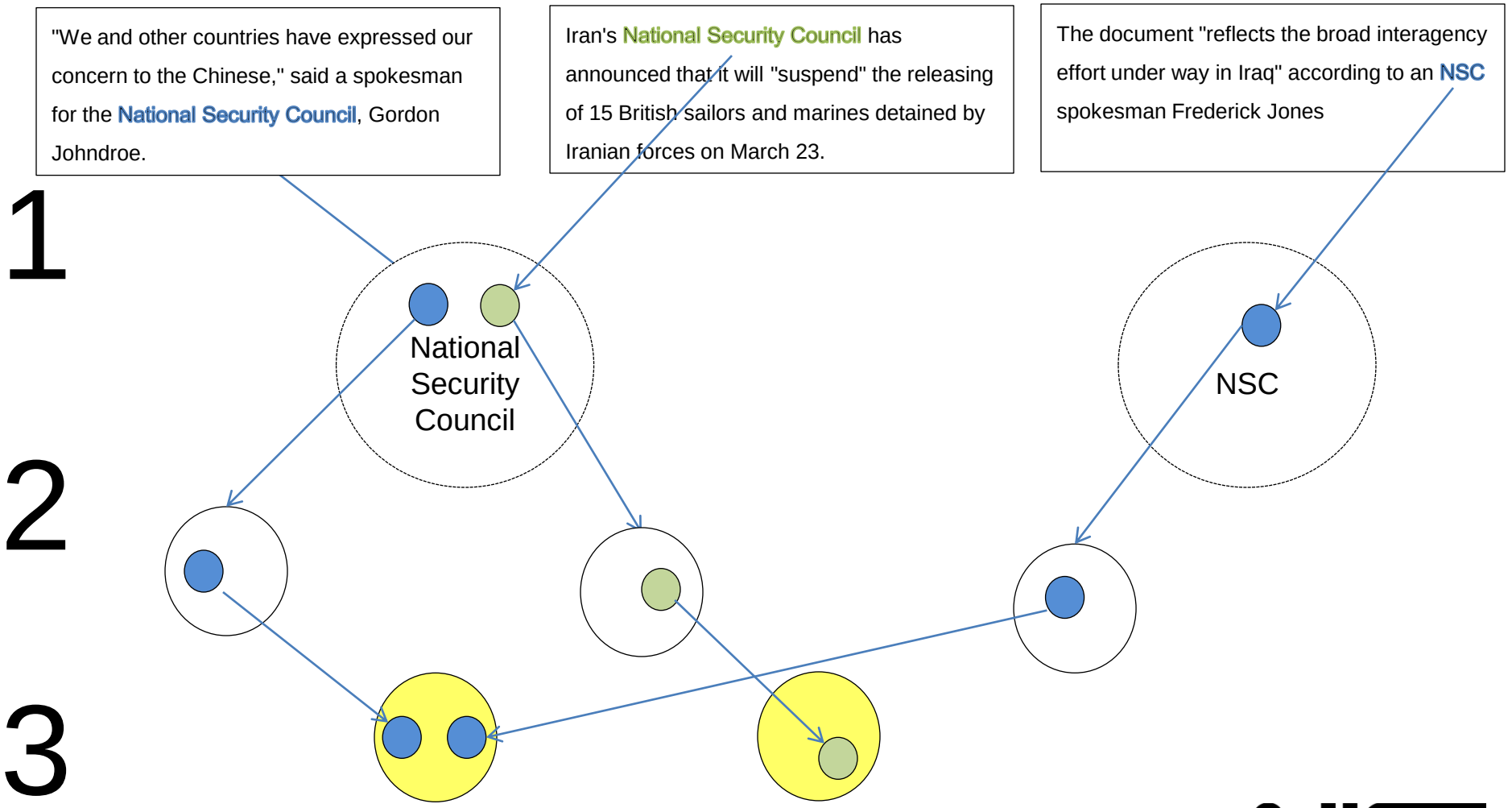
- Calculate similarity between mentions with a logistic regression classifier
  - (Mayfield et al., 2009)

Key Features

Feature Category	Description
Entity Type	Person, organization, etc…
Entity Links	Existence and confidence of same KB identifier (non-clustering)
Term Similarity	TFIDF weighted bag of words (Bagga/Baldwin 1998)
Local Context	E.g.: Actor Will Smith or Vice-President Will Smith

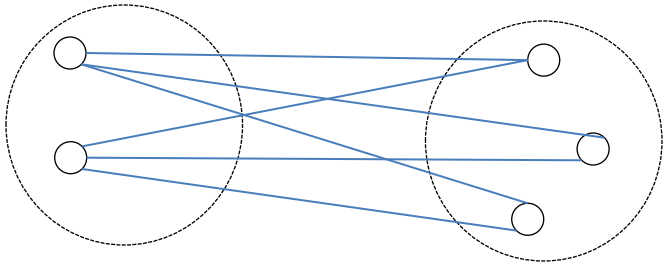
CDC: Stage 3

Merge across clusters



CDC: Stage 3 Model

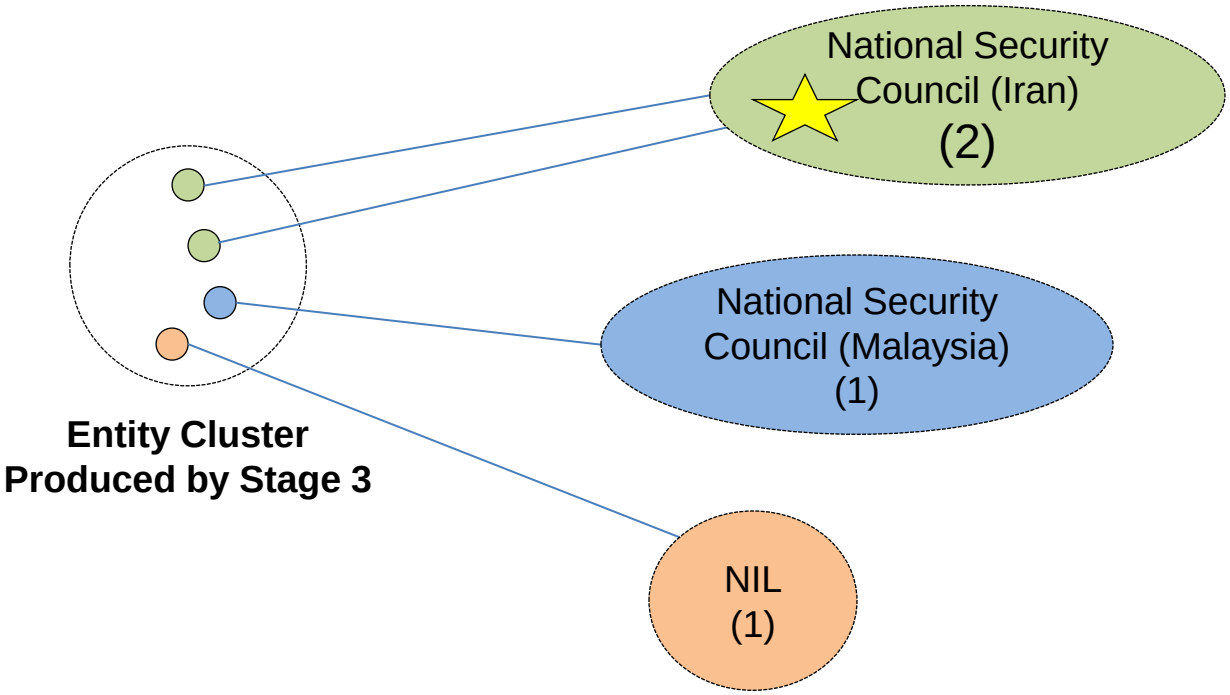
Function	Description
$I_1 = 1$	If m_1 and m_2 have same KB identifier w/ confidence $> \mu$
$I_2 = 1$	If m_1 and m_2 are embedded in a longer common phrase



$$\sum_{m_1 \in M_1} \sum_{m_2 \in M_2} \alpha_k I_k(m_1, m_2) > \lambda, k \in (1,2, \dots)$$

Stage 4: KB Identifier Generation

- Map each cluster to the knowledge base.
- Voting algorithm
  - Each entity link has a weight of 1



- Experimented with weighted links

English Entity Linking Submission

- 3 submissions
  - LCC3: Entity Linking with NIL Clustering System, without web access
    - Primary Evaluation
  - LCC1: Same as LCC3, with web access
  - LCC2: Changed model parameters to target precision

Submission	P	R	F
LCC3*	84.4	84.7	**84.6**
LCC1	86.7	87.1	86.9
LCC2	86.7	86.2	86.4

2011 KBP Submissions

- Attempting to improve precision ended up hurting recall

Inductive vs. Deductive Experiments

- Inductive System
  - Non-Clustering Linking as a feature
- Deductive System
  - Non-Clustering Linking as ground truth

System	P	R	F	MicroAvg
Inductive	84.4	84.7	84.6	86.1
Deductive	84.2	83.7	84.0	85.7

2011 Eval Set

- +0.6 F
- +0.4 MicroAvg

Use of Non-Clustering Entity Linking Features

- Inductive system
  - Entity Links as a feature in Stages 2 and 3
  - Entity Links used to assign KB in Stage 4
- Without links as cluster features
  - Only uses entity links in Stage 4

System	P	R	F	MicroAvg
Inductive	84.4	84.7	84.6	86.1
without links	82.1	83.2	82.7	84.7

2011 Eval Set

- +1.9 F
- +1.4 MicroAvg

2011 Non-Clustering Entity Linking Improvements

- Utilize Local Context
  - "Jim moved from Missouri to **Springfield**, Illinois."
  - "Joe lives in Atlanta, **Georgia**"
- String normalization (diacritics)
  - "Jose" → "José"
- More precise candidate generation

System	P	R	F	MicroAvg
2010	81.7	82.2	82.0	83.7
2011	84.4	84.7	84.6	86.1

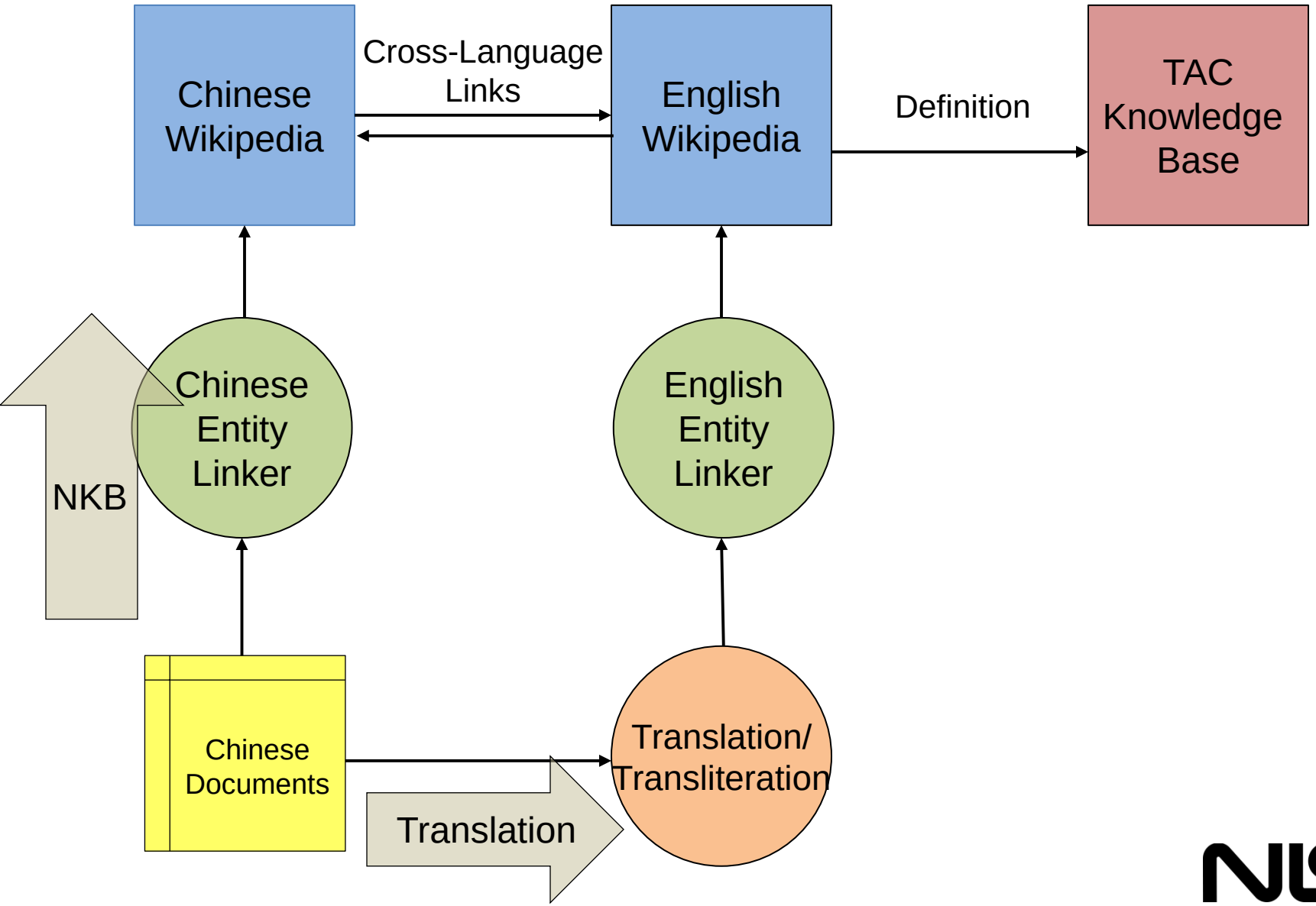
2011 Eval Set

- +2.6 F
- +2.4 MicroAvg

Talk Overview

1. Entity Linking with NIL Clustering
2. **Cross-lingual Entity Linking with NIL Clustering**
   - Why is this task important?
   - Added Challenges
     - Linking Chinese entities
     - Clustering Chinese entities
     - Clustering English and Chinese entities

Cross-Language Linking Approaches



Native Language Knowledge Base Approach

- Link to the *Native Language Knowledge Base (NKB)*
- Wikipedia provides a useful knowledge base in many languages
  - 39 languages with > 100k pages
- Adapting our system to go from English to Chinese
  - See (Lehmann et al., 2010)
  - Candidate Generation
    - Wikipedia-based sources apply equally
    - Sources like acronym do not work
    - Search engine: “site:zh.wikipedia.org”
  - Candidate Ranking
    - Using low ambiguity link similarity
  - NIL Detection
    - Trained model for Chinese
  - Cluster Similarity
    - Context similarity using document context is language independent
    - Trained model for Chinese

Translation Approach

- Compared to NKB
  - Advantages: Can use our English linking system
  - Disadvantage: Translation fidelity
  - Unknown: Chinese vs. English entities
- Translate the query documents and queries (using Bing Translation API)
  - Use English system directly
- NKB performs 1.9 F better
- Combination algorithm
  - Run both systems, select most confident link, prefer non-NIL over NIL

System	F
NKB	80.9
Translation	79.0
Voting	82.6

Score on Development Set

  - +1.7 F

Cross-Lingual Scores

- 3 submissions
  - LCC1: NKB (no web)
    - * Primary Evaluation
  - LCC2: NKB (with web)
  - LCC3: NKB (with web) combined Translation

Submission	P	R	F	Gain (F)
LCC1*	78.6	79.0	**78.8**	
LCC2	80.7	81.2	80.9	+2.1
LCC3	78.8	81.3	80.0	+1.2

2011 KBP Cross-Lingual submissions

- +2.1 F with Web Features
- +1.2 F with Combined

Chinese vs. English linking

- Cross-lingual data contains both English and Chinese queries

Submission	Combined	English	Chinese
LCC1 (no web)	82.4	84.6	81.3
LCC2	84.3	87.3	82.9
LCC3	83.9	87.5	82.2

Entity Linking Scores by language

- English several % better
- +1.6 F with Chinese Web

Development vs. Evaluation

- In development set, the combination system performed better than NKB system

System	Dev Set	Eval Set	Gain
NKB	80.9	**82.9**	+2.0
Translation	79.0	79.8	+0.8
Voting	82.6	82.2	-0.4

Entity Linking Scores (dev vs. eval)

- Both NKB and Translation performed better on evaluation set

Conclusions

- Inductive outperforms Deductive
- NKB outperforms Translation
  - Combined approach promising
- Clustering and Linking require little language customization
  - Could be an area for improvements
- Currently addressing scalability
  - Built a distributed clustering algorithm
    - Stores result in NoSQL database
    - Web front end
  - Working to scale to millions to documents
    - (Singh et al., 2011)

- Thank You!