WIKIFICATION

Heng Ji

jih@rpi.edu

April 16, 2014
Outline

- Background
- A Skeletal View
- Key Challenges and Recent Advances
- New Tasks, Trends and Applications
Limitations of Traditional IE/QA Tracks

- Traditional Information Extraction (IE) evaluations (e.g. Message Understanding Conference (MUC) / Automatic Content Extraction (ACE) programs)
  - Assess the ability to extract information from individual documents in isolation
  - In practice, we may need to gather information about a person or organization that is scattered among the documents of a large collection
- New aspects of KBP
  - Identify the relevant documents and integrate facts, possibly redundant, possibly complementary, possibly in conflict, coming from multiple documents
  - Use the extracted information to augment an existing data base

- Traditional Question Answering (QA) Evaluations
  - Limited efforts on disambiguating entities in queries (e.g. Pizzato et al., 2006)
  - Limited use of relation/event extraction in answer search (e.g. McNamee et al., 2008)
The Goal of KBP

• General Goal
  • Bridge the IE and QA communities to promote research in discovering facts about entities and expanding a knowledge source
  • Conducted as part of the NIST Text Analysis Conference
    • 23 teams submitted results for one or both sub-tasks in KBP2010

• What’s New
  • Extraction at large scale (1.3 million documents);
  • Using a representative collection (not selected for relevance);
  • Cross-document entity resolution (extending the limited effort in ACE);
  • Linking the facts in text to a knowledge base;
  • Distant (and noisy) supervision through Infoboxes;
  • Rapid adaptation to new relations;
  • Support multi-lingual information fusion (KBP2011);
  • Capture temporal information (KBP2011)
Steve Jobs, Apple founder, dies

October 05, 2011 | By Brandon Griggs, CNN

Steve Jobs, the visionary in the black turtleneck who co-founded Apple in a Silicon Valley garage, built it into the world’s leading tech company and led a mobile-computing revolution with wildly popular devices such as the iPhone, died Wednesday. He was 56.

The hard-driving executive pioneered the concept of the personal computer and of navigating them by clicking onscreen images with a mouse. In more recent years, he introduced the iPod portable music player, the iPhone and the iPad tablet -- all of which changed how we consume content in the digital age.
Overview of KBP Tasks

- Documents in T
  - Mono-lingual Entity Linking
  - Mono-lingual Slot Filling
  - Temporal Slot Filling
  - Temporal Tuples
  - Temporal Base in T

- Documents in S
  - Cross-lingual Entity Linking
  - Cross-lingual Slot Filling
we need some **bass**!

what do you *mean*...?
no.

here you go!
Still no.

Ah... of course...
That’s actually quite interesting! Still no.

Take your pick!

(from wordnet.princeton.edu)

**Noun**

- **S:** (n) **bass** (the lowest part of the musical range)
- **S:** (n) **bass, bass part** (the lowest part in polyphonic music)
- **S:** (n) **bass, basso** (an adult male singer with the lowest voice)
- **S:** (n) **sea bass, bass** (the lean flesh of a saltwater fish of the family Serranidae)
- **S:** (n) **freshwater bass, bass** (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
- **S:** (n) **bass, bass voice, basso** (the lowest adult male singing voice)
- **S:** (n) **bass** (the member with the lowest range of a family of musical instruments)
- **S:** (n) **bass** (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

**Adjective**

- **S:** (adj) **bass, deep** (having or denoting a low vocal or instrumental range) “a deep voice”; “a bass voice is lower than a baritone voice”; “a bass clarinet”
BINGO

I could have used some context =( 
At a WH briefing here in Santiago, NSA spox Rhodes came with a litany of pushback on idea WH didn't consult with.
Entity Linking: Create Wiki Entry?

Shocking Jim Parsons truths revealed after Emmy win
August 20, 2010 | 5:22 pm

Jim Parsons played the kola kola bird in Rudyard Kipling's "The Elephant's Child," donning a pair of yellow lights

<query id="EL000304">
  <name>Jim Parsons</name>
  <docid>eng-NG-31-100578-11879229</docid>
</query>

- Query type: persons, GPEs, organizations
Disambiguation to Wikipedia (D2W)

- **Word Sense Disambiguation**
  - Link linguistic item \( q \) in a document \( Q \) to target \( d \) in domain \( D \)
  - \( D \subseteq \mathcal{P}(C) \)
  - \( C \) is a *sense inventory*

- **D2W:**
  - \( C := \{ \text{Wikipedia (WP) articles} \} \)
  - Often, \( D \subseteq \{ x \in \mathcal{P}(C) \mid |x|=1 \} \)
    - In this case, each \( d \) is one WP page
  - View output as:
    - \( \{ \{ (q, d) \} \forall q \in Q \} \forall Q \in \text{Corpus} \}
    - \( \{ (Q, \{ d \}) \forall Q \in \text{Corpus} \} \)
# Task Dimensions

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Examples</th>
</tr>
</thead>
</table>
| Restrictions on C          | ▪ Only consider Named Entities  
                            ▪ Consider any Wikipedia page a potential target                           |
| Choice of mentions to link | ▪ Provided in input → input = <document, list of mentions>  
                            ▪ Selected by system → input = document                                    |
| Mention overlap            | ▪ Omit sub-strings of linkable n-grams  
                            ▪ Link entire NP, N, and NP-internal modifier separately                      |
| Mode of Reference          | ▪ N-gram must denote concept directly  
                            ▪ Link concepts related to document topic                                      |
| Out-of-WP concepts         | ▪ Identify mentions that refer to an out-of-WP concept  
                            ▪ Cluster out-of-WP concept mentions by referent                                |
| Number of WP Articles per q | ▪ Link each N-gram to 0 or 1 concepts  
                            ▪ Link each N-gram to 0 or more concepts                                       |
Evaluation of D2W

**in vitro**

1. Text
2. D2W
3. Evaluate
4. Results

5. GOLD STD D2W

**in vivo**

1. Text
2. D2W
3. Other NLP Task
4. Evaluate
5. Results
6. GOLD STD NLP TASK

Further Discussion: (Navigli, 2009)
D2W Downstream – recent work

• Co-reference resolution (Ratinov & Roth, 2012)
  • “After the vessel suffered a catastrophic torpedo detonation, Kursk sank in the waters of Barents Sea…”
  • Knowing Kursk → *Russian submarine K-141 Kursk* helps system to co-ref “Kursk” and “vessel”

• Document classification (Vitale et. al., 2012)
  • Tweets labeled *World, US, Science & Technology, Sports, Business, Health, Entertainment*
  • Features based only on topical representation in terms of WP concepts

• Information retrieval
  • Temporal web image retrieval (Dias et. al., 2012)
  • D2W yields *intent modifiers* that guide NL queries to logical representation based on entity type (Guerrisi et al 2012)
Data: Inter-Annotator Agreement for News

<table>
<thead>
<tr>
<th>Entity Type</th>
<th>#Total Queries</th>
<th>Agreement Rate</th>
<th>Genre</th>
<th>#Disagreed Queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>59</td>
<td>91.53%</td>
<td>Newswire</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Web Text</td>
<td>1</td>
</tr>
<tr>
<td>Geo-political</td>
<td>64</td>
<td>87.5%</td>
<td>Newswire</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Web Text</td>
<td>5</td>
</tr>
<tr>
<td>Organization</td>
<td>57</td>
<td>92.98%</td>
<td>Newswire</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Web Text</td>
<td>1</td>
</tr>
</tbody>
</table>
For Tweets: False Positives Deemed Correct

<table>
<thead>
<tr>
<th>Type</th>
<th>FPDC</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>New</td>
<td>So excited to announce I’ll be singing &quot;God Bless America&quot; during the 7th Inning Stretch at the Detroit Tigers... URL</td>
<td>#NP &quot;Crazy&quot;- The Boys - <em>heeeey</em></td>
</tr>
<tr>
<td>Replace</td>
<td>Sweet 16! What a good feeling. Keep it going... Go Gators!!!</td>
<td><em>Sweet sixteen (birthday) → NCAA Men’s Division I Basketball Championship</em></td>
</tr>
<tr>
<td>Eq.</td>
<td>The Devil is a liar! Thank God for giving you to chance to see this beautiful morning. I’m thankful and very blessed.</td>
<td>Satan was deemed conceptually equivalent to Devil</td>
</tr>
<tr>
<td>Hash</td>
<td>#NATO to enforce arms embargo against #Libya - URL #Gaddafi</td>
<td>Named entities are linked throughout the dataset.</td>
</tr>
</tbody>
</table>

- Corrected when...
  - Analogous annotation found in the gold standard
  - Gold standard deemed unambiguously incorrect
  - GLOW’s target deemed analogous to gold standard
  - Hash tags were deemed permissible

70 additional annotations
...and “partially correct”

<table>
<thead>
<tr>
<th>Type</th>
<th>FPD <em>partially</em> C</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part of linked term</td>
<td>I am rooting for the Wash U Bears in this week's NCAA Div III Final Four. Go, ladies! #fgs</td>
<td>Washington University in St. Louis is too vague compared to Washington University Bears</td>
</tr>
<tr>
<td>Part of linked term</td>
<td>Don't u hate when u go to sleep then wake up with a stuffy nose? annoyed</td>
<td>Stuffy nose → Rhinitis is good. “nose” → nose violates longest n-gram heur.</td>
</tr>
<tr>
<td>Too general</td>
<td>Photos are great for engaging with your audiences. Upload images to Flickr.com and create slideshows with URL</td>
<td>Uploading and downloading too vague. Photo Sharing not referenced directly?</td>
</tr>
<tr>
<td>Too general</td>
<td>Section of I-55 Closed Until Monday: I-55 will be closed in both directions between Carondelet and...</td>
<td>Interstate 55 in Missouri is better than Interstate 55 alone; however, would be Interstate 55 in Louisiana</td>
</tr>
<tr>
<td>Not a central topic</td>
<td>God Bless Japan</td>
<td>God omitted, Japan included, since tweet is about Japan.</td>
</tr>
<tr>
<td>Correct sense, out of context</td>
<td>Waiting on my doorbell to ring for my Christmas feast! Sweet N sour chicken, vegetable fried rice ...</td>
<td>Chicken (food) is good, but Chicken is better than Chicken (game); (dance); (film); (city); (coward)</td>
</tr>
</tbody>
</table>

- Some mistakes are more forgivable
- Good mistakes may be better in vivo than bad mistakes
- Partially correct if sense was correct out of context
D2W \textit{in vitro} performance

- **Selection of mentions**
  - Does the system choose to link the terms chosen by annotators for linking?

- **Ranking accuracy**
  - For mentions that were linked by annotators, does the system provide the same output?

- **Linking accuracy**
  - Does the system provide the correct response, including out-of-WP, for mentions whose surface form was ever linked?

- **NIL clustering**
  - Does the system accurately cluster out-of-WP concepts by referent?

- **Other (including IR-inspired) metrics**
  - E.g. MRR, MAP, R-Precision, Recall, accuracy
Scoring: B-cubed++

- We use a tuple 
  \( \langle \text{doc-id, start, end, entity-type, kb-id} \rangle \)
  to represent each entity mention, where a special type of kb-id is \( NIL \).
Scoring

- Let $s$ be an entity mention in the system output, $g$ be an corresponding gold-standard. An output mention $s$ matches a reference mention $g$ iff:

  1. $s.doc-id = g.doc-id$,
  2. $s.start = g.start$, $s.end = g.end$,
  3. $s.entity-type = g.entity-type$,
  4. $s.kb-id = g.kb-id$. 


Scoring

We use $C_g(s_i)$ to denote the cluster in $g$ that contains the mention $s_i$. If no mention in $g$ can match $s_i$, then $C_g(s_i) = \emptyset$. Likewise, $C_s(g_j)$ denotes the cluster in $s$ that contains the mention $g_j$. Then the recall of B-Cubed++ score is calculated as follows:

$$\text{recall} = \frac{\sum_{s_i \in s, C_g(s_i) \neq \emptyset} \frac{|C_s(s_i) \cap C_g(s_i)|}{|C_g(s_i)|}}{|g|}$$
The numerator counts the ratio of each correct mention in its reference cluster, and the final recall is averaged by the total number of mentions in the gold-standard. Similarly, the precision of B-Cubed++ score is calculated as follows:

\[
\text{precision} = \frac{\sum_{g_j \in g, C_s(g_j) \neq \emptyset} \frac{|C_g(g_j) \cap C_s(g_j)|}{|C_s(g_j)|}}{|s|}
\]
B³++: Precision

- Counting: intersecting each system output w/ gold standard
- A mention’s credit: ratio of correct system output mentions
- Precision = sum mention credits / #system-output-mentions
  \[ \frac{(1/2 + 0 + 0 + 1/1 + 0)}{6} = \frac{1}{4} \]
B³++: Recall

- Counting: intersecting each gold standard with system output
- A mention’s credit: ratio of correct gold standard mentions
- Recall = sum mention credits / #gold-standard-mentions
  = \((1/3 + 0 + 0 + 1/2 + 0)/6 = 5/36\)
\[ B^3++: \text{Precision (NIL case)} \]

- Precision = \( \frac{\text{sum mention credits}}{\#\text{system-output-mentions}} \)
  \[ = \frac{1/1 + 2/2 + 2/2 + 2/2 + 2/2 + 1/1}{6} = 1 \]
$B^3++$: Recall (NIL case)

- Recall = sum mention credits / #gold-standard-mentions
  
  $$= \frac{1/3 + 2/3 + 2/3 + 2/2 + 2/2 + 1/1}{6} = 0.78$$
Outline

• Background
• A Skeletal View
• Key Challenges and Recent Advances
• New Tasks, Trends and Applications
Commonness Baseline

\[ P(q \Rightarrow c) = \frac{\text{count}(q \rightarrow c)}{\sum_{c' \in C} \text{count}(q \rightarrow c')} \]

⇒ := Link from corpus to Wikipedia
→ := Link from within Wikipedia

The probability that an n-gram \( q \) in a tweet \( Q \) refers to a concept \( c \) is equal to the number of times \( q \) serves as a hyperlink, \textbf{within Wikipedia}, to \( c \), divided by the number of times \( q \) serves as a hyperlink, \textbf{within Wikipedia}, to any concept \( c' \).
Commonness baseline

\[
P(q \rightarrow c) := \frac{\text{count}(q \rightarrow c \text{ within } WP)}{\sum_{c' \in W} \text{count}(q \rightarrow c')}
\]

- **Recall** means the proportion of correct concepts, regardless of ranking.
- **Proportion of mentions that are “solvable” via re-ranking.**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>60.21%</td>
</tr>
<tr>
<td>R-Prec</td>
<td>52.71%</td>
</tr>
<tr>
<td>Recall*</td>
<td>77.75%</td>
</tr>
<tr>
<td>MRR</td>
<td>70.80%</td>
</tr>
<tr>
<td>MAP</td>
<td>58.53%</td>
</tr>
</tbody>
</table>

**Compare with:**

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Cmns Rec**</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE</td>
<td>86.85%</td>
</tr>
<tr>
<td>MSNBC</td>
<td>88.67%</td>
</tr>
<tr>
<td>AQUAINT</td>
<td>97.83%</td>
</tr>
<tr>
<td>Wiki</td>
<td>98.59%</td>
</tr>
</tbody>
</table>

- **Commonness is simple, yet effective.**
- **But performance degrades with corpus “distance” from Wikipedia.**

*Based on the top 50 concepts returned for a tweet
*Based on the top 20 concepts returned for an n-gram
Enhancing Commonness

For an n-gram $q$ in a document $Q$, supervised re-ranking of commonness output

Meij et al. (2012)
- Only 5/33 features are in terms of entire $Q$ and candidate concepts for $q$
- Other features in terms of $q$, $c$, $(q, c)$
- But **not the candidates concepts for other $q'$ in $Q$**

GLOW – Ratinov et al (2011)
- Whether $q \rightarrow c$ informed by $Rel(c, c')$
- For other $q'$ in $Q$, such that $q' \rightarrow c'$ is likely
- In tweets there are few helpful $q'$ ... more context is needed!

**Goal**
- Augment $Q$ with potentially informative tweets, $Q'$
- Use a system that takes **the right amount** surrounding context into account by maximizing **global coherence** of a **disambiguation context**
GLOW (Ratinov et al., 2011)

\[
P(q \rightarrow c) = \frac{\text{count}(q \rightarrow c \text{ within WP})}{\sum_{c' \in W} \text{count}(q \rightarrow c')}
\]

\[
\phi_1(q, c) = P(c)
\]

\[
\phi_2(q, c) = P(q \rightarrow c)
\]

\[
\phi_3(q, c) = \cos\_sim(\text{Text}(q), \text{Context}(c))
\]

\[
\phi_4(q, c) = \cos\_sim(\text{Text}(q), \text{Text}(c))
\]

\[
\psi_1(c_i, c_j) = I_{[c_i \rightarrow c_j]} \text{NGD(In(c_i), In(c_j))}
\]

\[
\psi_1(c_i, c_j) = I_{[c_i \rightarrow c_j]} \text{PMI(Out(c_i), Out(c_j))}
\]

\[
\Gamma^* = \arg\max_{\Gamma} \sum_{i=1}^{n} \phi(q_i, c_{i}^{top}) \sum_{c_j^{top} \in \Gamma} \psi(c_i^{top}, c_j^{top})
\]
Candidate Concepts for $q'$
GLOW at a glance

- Retrieve candidates with high recall method

- Re-rank

- Choose Top Result for each N-gram

- For each $q$, retrieve all $c$ such that $\text{Commonness}(q, c) > 0$
GLOW (Ratinov et al., 2011)

\[ P(q \rightarrow c) := \frac{\text{count}(q \rightarrow c \text{ within } WP)}{\sum_{c' \in W} \text{count}(q \rightarrow c')} \]

\[ \phi_1(q, c) = P(c) \]
\[ \phi_2(q, c) = P(q \rightarrow c) \]
\[ \phi_3(q, c) = \cos\text{ _sim(Text}(q), \text{Context}(c)) \]
\[ \phi_4(q, c) = \cos\text{ _sim(Text}(q), \text{Text}(c)) \]
\[ \ldots \]

\[ \psi_1(c_i, c_j) = I_{[c_i \rightarrow c_j]} \text{NGD}(\text{In}(c_i), \text{In}(c_j)) \]
\[ \ldots \psi_1(c_i, c_j) = I_{[c_i \rightarrow c_j]} \text{PMI}(\text{Out}(c_i), \text{Out}(c_j)) \]

\[ \Gamma^* = \arg\max_{\Gamma} \sum_{i=1}^{n} \phi(q_i, c_i^{\text{top}}) \sum_{c_j^{\text{top}} \in \Gamma} \psi(c_i^{\text{top}}, c_j^{\text{top}}) \]
General Architecture

- **Statistical Name Variant Expansion (NUSchime)**
  - “CCP” vs. “Communist Party of China”
  - “MINDEF” vs. “Ministry of Defence”

- **New Ranking Algorithms**
  - e.g. ListNet (CUNY), Random Forests (THUNLP, DMIR_INESCID)

- **Query Classification**
  - DMIR_INESCID, CUNY, MSRA

- **Go Beyond Single Query and Single KB Entry**
  - Wikification (UIUC), Collaborative ranking (CUNY), Link all entities and inference (MS_MLI, CMCRC)
Basic Methods

- Prior Popularity
  - Assume the most prominent entity for a given mention is the most probable underlying concept for that mention.

- Context Similarity
  - Define a similarity measure between the text around the concept mention and the document describing the referent concept in the knowledge base.

- Topical Coherence
  - Define a topical/semantic coherence measure between the mention’s referent concept and other mentions within the same context.

- Mostly assume mentions are given
Ranking Approach Comparison

- Unsupervised or weakly-supervised learning
  - annotated data is minimally used to tune thresholds and parameters
  - The similarity measure is largely based on the unlabeled contexts

- Supervised learning
  - a pair of entity and KB node is modeled as an instance for classification
  - Such a classifier can be learned from the annotated training data based on a wide variety of features

- Graph-based ranking
  - context entities are taken into account in order to reach a global optimized solution together with the query entity

- IR approach
  - the entire source document is considered as a single query to retrieve the most relevant Wikipedia article
## Typical Ranking Features

<table>
<thead>
<tr>
<th>Feature Category</th>
<th>Feature Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Name</strong></td>
<td><strong>Spelling match</strong> Exact string match, acronym match, alias match, string matching…</td>
</tr>
<tr>
<td><strong>KB link mining</strong></td>
<td>Name pairs mined from KB text redirect and disambiguation pages</td>
</tr>
<tr>
<td><strong>Name Gazetteer</strong></td>
<td>Organization and geo-political entity abbreviation gazetteers</td>
</tr>
<tr>
<td><strong>Document surface</strong></td>
<td><strong>Lexical</strong> Words in KB facts, KB text, query name, query text.</td>
</tr>
<tr>
<td></td>
<td>Tf.idf of words and ngrams</td>
</tr>
<tr>
<td><strong>Position</strong></td>
<td>Query name appears early in KB text</td>
</tr>
<tr>
<td><strong>Genre</strong></td>
<td>Genre of the query text (newswire, blog, …)</td>
</tr>
<tr>
<td><strong>Local Context</strong></td>
<td>Lexical and part-of-speech tags of context words</td>
</tr>
<tr>
<td><strong>Entity Context</strong></td>
<td><strong>Type</strong> Query entity type, subtype</td>
</tr>
<tr>
<td></td>
<td><strong>Relation</strong> Entities co-occurred, attributes/relations/events with the query</td>
</tr>
<tr>
<td></td>
<td><strong>Coreference</strong> Coreference links between the source document and the KB text</td>
</tr>
<tr>
<td><strong>Profile</strong></td>
<td>Slot fills of the query, KB attributes</td>
</tr>
<tr>
<td><strong>Concept</strong></td>
<td>Ontology extracted from KB text</td>
</tr>
<tr>
<td><strong>Topic</strong></td>
<td>Topics (identity and lexical similarity) for the query text and KB text</td>
</tr>
<tr>
<td><strong>KB Link Mining</strong></td>
<td>Attributes extracted from hyperlink graphs of the KB text</td>
</tr>
<tr>
<td><strong>Popularity</strong></td>
<td><strong>Web</strong> Top KB text ranked by search engine and its length</td>
</tr>
<tr>
<td></td>
<td><strong>Frequency</strong> Frequency in KB texts</td>
</tr>
</tbody>
</table>
What Works: Entity Profiling

**Disambiguation**

**Name Variant Clustering**

9.3-14.3% absolute improvement (BuptPris, CUNY, HLTCOE)
What Works: Modeling Coherence

- CUNY team: Collaborative Ranking (Chen and Ji, EMNLP2011)
  - Automatic profiling for each entity
  - Construct a collaborative network for the target entity based on graph-based clustering
  - Rank multiple decisions from collaborative entities (micro) and algorithms (macro)
  - 7% absolute improvement; performance comparable to the top 2 system in KBP2010
- WebTLab team and CMCRC team disambiguated all entities in the contexts with PageRank
- SMU-SIS team re-formulated queries using contexts
- LCC team modeled contexts using Wikipedia page concepts, and computed linkability scores iteratively

Correct rank: $A < B$

$$f(q, o_A^{(q)}) = 0.7, f(q, o_B^{(q)}) = 0.3, A > B$$

$$g_1(A) = 0.4, g_2(B) = 0.6, A < B$$
Unsupervised/Minimally-Supervised vs. Supervised Learning for Entity Linking
Impact of Semantic Features on Entity Linking

- **CUNY-BLENDER**: Use Slot Filling results as features (entity profile)

<table>
<thead>
<tr>
<th>System</th>
<th>Person</th>
<th>Organization</th>
<th>Geo-Political</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without SF feedback</td>
<td>84.6%</td>
<td>63.1%</td>
<td>57.5%</td>
<td>59.9%</td>
</tr>
<tr>
<td>With SF feedback</td>
<td>92.8%</td>
<td>65.7%</td>
<td>84.1%</td>
<td>69.3%</td>
</tr>
</tbody>
</table>

- **BuptPris**: Use name tagging, infoboxes etc. as features

<table>
<thead>
<tr>
<th>System</th>
<th>Person</th>
<th>Organization</th>
<th>Geo-Political</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without semantic Features</td>
<td>83.9%</td>
<td>59.5%</td>
<td>33.4%</td>
<td>58.9%</td>
</tr>
<tr>
<td>With semantic features</td>
<td>79.1%</td>
<td>74.1%</td>
<td>66.6%</td>
<td>73.3%</td>
</tr>
</tbody>
</table>
Impact of Data Genre (All queries)
Impact of Data Genre (Non-NIL queries)
Outline

• Background
• A Skeletal View
• **Key Challenges and Recent Advances**
• New Tasks, Trends and Applications
Analysis of Difficult Mentions

• There are 2250 queries in the Entity Linking evaluation
  • for 58 of them at most 5 (of 46) systems produced correct answers
  • most of these queries have corresponding KB entries
  • for 19 queries no system found the correct answer
• 11 queries are highly ambiguous city names which can exist in many states or countries (e.g. “Chester”), or can refer to person or organization entities
  • Contextual clues may be quite subtle:

  ...and was from a Mission School in Canton, ... but for the energetic efforts of this Chinese Christian man and the Refuge Matron...

  in China
Challenge 1: Background Knowledge

PROXY_AFP_ENG_20050604_0444.1 2005-06-05
PROXY_AFP_ENG_20050604_0444.2 Taiwan (TW)
PROXY_AFP_ENG_20050604_0444.3 International; weapons
PROXY_AFP_ENG_20050604_0444.4 Taiwan successfully fired its first cruise missile.
PROXY_AFP_ENG_20050604_0444.5 This will enable Taiwan to hit major military targets in southeast China.
PROXY_AFP_ENG_20050604_0444.6 050605 the China Times reported that Taiwan has successfully test fired the Hsiung Feng its first cruise missile enabling Taiwan to hit major military targets in southeast China.

Hsiung Feng IIE was deployed in 2005
News reporting date is 2005/6/5

Linking to “Hsiung Feng IIE” model
During talks in Geneva attended by US Undersecretary of State William J. Burns Iran refused to respond to Solana’s offers.

Wiki title candidates:
William_J._Burns (1861-1932)
William_Joseph_Burns (1956- )

William_J._Burns was dead in 1932, so linking to William_Joseph_Burns
Challenge 3: Context Entities

•DF-183-195681-794_9722.3 Hundreds of protesters from various groups converged on the state capitol in Topeka, Kansas today.

•DF-183-195681-794_9722.8 Second, I have a really hard time believing that there were any ACTUAL “explosives” since the news story they link to talks about one guy getting arrested for THREATENING Governor Brownback.

•Linking Brownback to Sam Brownback who is the governor of Kansas
Challenge 4: Collective Inference

• DF-170-181104-872_8783.1 Where would McCain be without Sarah?

• McCain = John_McCain
• Sarah = Sarah_Palin

• Sarah was the nominee for Vice President in the 2008 presidential election alongside John McCain
Challenge 5: Deep Relations

• The Yuri dolgoruky is the first in a series of new nuclear submarines to be commissioned this year but the bulava nuclear-armed missile developed to equip the submarine has failed tests and the deployment prospects are uncertain.
• The Yuri dolgoruky is the first in a series of new nuclear submarines to be commissioned this year but the bulava nuclear-armed missile developed to equip the submarine has failed tests and the deployment prospects are uncertain.
Challenges for Wikification for Social Media

• Messages are **short, noisy and informal**
  • Lack of rich context to compute context similarity and ensure topical coherence

• Lack of **Labeled Data** for Supervised Model
  • Lack of Context makes annotation more challenging
  • Need to search more background information

• Better Models:
  • performing **collective inference**
  • demanding **less labeled data**
Deeper Analysis

• Mentions and their correct referent concepts tend to share a set of characteristics
  • String similarity (&lt;Chicago, Chicago&gt; and &lt;Facebook, Facebook&gt;)

• Two coreferential mentions should be linked to the same concept
  • User A: At a WH briefing here in Santiago, NSA spox Rhodes…
  • User A: Chinese president is going to visit White House and meet Obama….
Deeper Analysis (cont’)

- Two highly semantically-related mentions are more likely to be linked to two highly semantically-related concepts

  Stay up **Hawk Fans**. We are going through a **slump**, but we have to stay positive. Go **Hawks**!
Advanced Approach (Huang et al., 2014)

- Semi-Supervised Graph Regularization
  - Relational Graph with *heterogeneous* relations
    - Local Compatibility
    - Coreference
    - Semantic Relatedness

- Advantages:
  - Incorporate global evidence from *multiple messages* with fine-grained relations
  - Perform *collective inference* to identify and link a set of relevant mentions
  - Make use of manifold (cluster) structure and need less training data
  - Collective inference over multiple messages help solve information shortage problem
Semi-Supervised Graph Regularization (Zhu et al. 2003)

\[ Q(\mathcal{Y}) = \mu \sum_{i=l+1}^{n} (y_i - y_i^0)^2 + \frac{1}{2} \sum_{i,j} W_{ij} (y_i - y_j)^2. \]

- **Loss Function**: ensure the refined labels is not too far from the initial labels
- **Regularizer**: smooth the refined labels over the constructed graph
- Both closed and iterative form solutions exist
Relational Graph

- Each pair of mention m and concept c as a node
  - m is linkable, and c is the correct concept, \(<m, c>\) should be assigned label 1, otherwise 0

An example of the relational graph
Meta Path

• A meta-path is a path defined over a network and composed of a sequence of relations between different object types (Sun et al., 2011)
• Meta paths between mention and mention
  • M-T-M
  • M-T-U-T-M-M
  • M-T-H-T-M
  • M-T-U-T-M-T-H-T-M
  • M-T-H-T-M-T-U-T-M

M: mention, T: tweet, U: user,

Schema of a Heterogeneous Information Network in Informal Genre
Relational Graph Construction

- **Local Compatibility**
  - A set of local features based on mention, concept, mention + concept, and context

- **Coreference**
  - meta path

- **Semantic Relatedness**
  - meta path and link structure in Wikipedia

- Linear Combination of the three relations
Data and Scoring Metric

- **Data**
  - A public data set includes 502 messages from 28 users
  - A Wikipedia dump on May 3, 2013

- **Scoring Metric**
  - Standard precision, recall and F1
  - A pair of mention and concept is judged as correct
    - Mention is linkable
    - Concept is the correct referent concept
Models for Comparison

• TagMe: an unsupervised model based on prior popularity and semantic relatedness of a single message (Ferragina and Scaiella, 2010)
• Meij: the state-of-the-art supervised approach based on the random forest model (Meij et al., 2012)
• SSRegu: our proposed semi-supervised graph regularization model
Overall Performance (Huang, et al, ACL2014)

- **Meij**: use 100% labeled data
- **SSRegu**: use 50% labeled data

5% absolute F1 gain over the state-of-the-art supervised models
Our model with 30% labeled data achieves similar performance with the state-of-the-art supervised model.
Outline

• Background
• A Skeletal View
• Key Challenges and Recent Advances
• New Tasks, Trends and Applications
Cross-lingual Entity Linking

<query id="SF114">
  <name>李安</name>
  <docid>XIN20030616.0130.0053</docid>
</query>

Parent: Li Sheng
Birth-place: Taiwan Pindong City
Residence: Hua Lian
Attended-School: NYU
From Mono-lingual to Cross-lingual

<table>
<thead>
<tr>
<th>Difficulty</th>
<th>Task</th>
<th>All</th>
<th>NIL</th>
<th>Non-NIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambiguity</td>
<td>Mono-lingual</td>
<td>12.9%</td>
<td>5.7%</td>
<td>9.3%</td>
</tr>
<tr>
<td></td>
<td>Cross-lingual</td>
<td>20.9%</td>
<td>14.0%</td>
<td>28.6%</td>
</tr>
</tbody>
</table>

B-cubed+ F-Measure
CLEL NIL Clustering Performance
‘Boring’ CLEL Pipeline A

- Strategy 1: MT + English EL
  - Essentially the same as monolingual entity linking if we assume that MT is flawless

- **Pro:**
  - Doesn’t require source language document annotations

- **Con:**
  - Relies on machine translation (MT) output
  - Suffers from MT and name translation errors
‘Boring’ Pipeline B

• Strategy 2: Link to Multilingual Knowledge Base
  • Link mention to wikipedia, use wikipedia structure to link to KB node

- Pro:
  - Doesn’t require full document translation
  - Can use training data in many languages
  - Human annotation (e.g. Interwiki links) is very reliable

- Con:
  - Relies heavily on the existence of an source language KB whose size is comparable to the target language KB
  - Not easily adaptable to other low-density languages
李娜

中央电视台收视率最高的19点《新闻联播》也创纪录的用将近5分钟的时间报道了李娜法网夺冠的消息。

2011-06-05 20:55:52

北京时间6月5日消息，在昨天李娜创纪录的在法网夺冠后，一时间她登上了国内外各大媒体的头版头条，成为今日人们谈论的焦点话题。就连中央电视台收视率最高的19点《新闻联播》也创纪录的用将近5分钟的时间报道了李娜法网夺冠的消息。

支持者的开场白这样说：“在4号进行的法网女单决赛中，中国网球名将李娜以2-0战胜了卫冕冠军、意大利人斯齐亚沃尼，成为大满贯赛事第一个捧起单打冠军奖杯的亚洲人。”随后画面切换到李娜昨天比赛时的场景，第一个镜头正是赛前她和斯齐亚沃尼隔着网相立合影留念的情形，这时候音响响起，画面上也随之转至昨天比赛的精彩镜头中，虽然三十岁的斯齐亚沃尼曾在去年的法网女单比赛中让李娜止步8强，但在今年的罗兰加洛斯红土场上，李娜一路过关斩将挺入决赛。首盘李娜凭借第五局的破发6-4取胜，第二盘，29岁的李娜更是首局就强力破发，随后保住自己的发球局后以2-0领先，但此后斯齐亚沃尼在第八局顽强的将比分扳平，并将比赛拖入抢七，在抢七中，李娜开局就连续抢分以6-0开局，在握有两个冠军点的情况下，李娜顶住压力，随着斯齐亚沃尼的回球出界，李娜以6-0拿下抢七，最终将自己的名字刻在了苏珊-朗格伦杯上。

随后，现场到李娜的采访，当现场主持人问到：“你今天看起来很放松，难道你一点都不紧张吗？”这位新科法网冠军洋溢着幸福的笑容，用一口流利的英文说道：“我当然紧张，但是我没有表现出来。”主持人随后惊讶的说道：“原来你是故意不表现出紧张啊。”李娜顽皮的笑了笑：“是啊，我的表情可是有欺骗性的。”主持人又说道希望明年再在这里看到她前来卫冕，李娜肯定的回答说：“当然，我明年肯定还会再回来的！”
Our Research Hypotheses

1. Query mentions can be disambiguated based on their “neighbors”, i.e. other entities with which they co-occur.

2. “One entity per topic cluster” – similar query mentions sharing topically-related contexts tend to link to the same KB entry.
Enhanced System (Cassidy et al., 2012)
How do we find our childhood friends?

- Entities are not mentioned with equal frequencies
- Entities cannot be translated with the same difficulties
- First clues – schools_attended (ORG), hometowns (GPE), social networks (PER), …
- These form an “information network” for each entity
- Modeling information networks (CUNY, DAI, …)
## Good Neighbors in Source Documents

<table>
<thead>
<tr>
<th>Context Types</th>
<th>Query</th>
<th>KB Node</th>
<th>Key Context</th>
<th>Context Sentence</th>
<th>Context Sentence Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start-Position Event</td>
<td>埃特尔 (Ertl)</td>
<td>NIL3</td>
<td>智利 (Chilean) 奥委会 (Olympic Committee) 选为 (elected) 主席 (chairman)</td>
<td>智利击剑联合会领导埃特尔今晚被选为该国奥委会新任主席</td>
<td>The leader of Chilean Fencing Federation Ertl was elected as the new chairman of this country’s Olympic Committee tonight.</td>
</tr>
<tr>
<td>Affiliation</td>
<td>国家医药局 (National Medicine s Agency)</td>
<td>NIL4</td>
<td>保加利亚 (Bulgarian)</td>
<td>保加利亚国家医药局</td>
<td>Bulgarian National Medicines Agency</td>
</tr>
<tr>
<td>Located Relation</td>
<td>精细化工厂 (Fine Chemical Plant)</td>
<td>NIL6</td>
<td>芜湖市 (Wuhu City)</td>
<td>芜湖市精细化工厂</td>
<td>Fine Chemical Plant in Wuhu City</td>
</tr>
<tr>
<td>Context Types</td>
<td>Query</td>
<td>KB Node</td>
<td>Key Context</td>
<td>Context Sentence</td>
<td>Context Sentence Translation</td>
</tr>
<tr>
<td>---------------</td>
<td>-------</td>
<td>---------</td>
<td>-------------</td>
<td>------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>Part-whole Relation</td>
<td>Fairmont</td>
<td>Fairmont, West Virginia</td>
<td>WV</td>
<td>Verizon coverage in WV is good along the interstates and in the major cities like Charleston, Clarksburg, Fairmont, Morgantown, Huntington, and Parkersburg.</td>
<td>-</td>
</tr>
<tr>
<td>部分-整体关系</td>
<td>曼彻斯特(Manchester)</td>
<td>Manchester, New Hampshire</td>
<td>新罕布什尔州(New Hampshire)</td>
<td>曼彻斯特(新罕布什尔州)</td>
<td>Manchester (New Hampshire)</td>
</tr>
<tr>
<td>Employer / Title</td>
<td>米尔顿(Milton)</td>
<td>NIL1</td>
<td>巴西(Brazil); 代表(representative)</td>
<td>巴西政府高级代表米尔顿</td>
<td>Milton, the senior representative of Brazil government</td>
</tr>
<tr>
<td></td>
<td>NIL2</td>
<td>厄瓜多尔皮钦查省(Pichincha Province, Ecuador); 省长(Governor)</td>
<td>厄瓜多尔皮钦查省省长米尔顿</td>
<td>Milton, the Governor of Pichincha Province, Ecuador</td>
<td></td>
</tr>
<tr>
<td><strong>Context Types</strong></td>
<td><strong>Query</strong></td>
<td><strong>KB Node</strong></td>
<td><strong>Key Context</strong></td>
<td><strong>Context Sentence</strong></td>
<td><strong>Context Sentence Translation</strong></td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------</td>
<td>-------------</td>
<td>----------------</td>
<td>----------------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td><strong>Co-occurrence</strong></td>
<td>塞维利亚 (Sevilla)</td>
<td>Sevilla, Spain</td>
<td>西班牙 (Spain)</td>
<td>西班牙两名飞行员15日举行婚礼，从而成为西班牙军队中首对结婚的同性情侣。婚礼在塞维利亚市政厅举行。</td>
<td>Two pilots had their wedding in <strong>Spain</strong> on 15th, and so they became the first homosexual couple who got married in Spanish troops. The wedding was held in <strong>Sevilla</strong> city hall.</td>
</tr>
<tr>
<td><strong>民主进步党 (Democratic Progressive Party)</strong></td>
<td>民主进步党 (Democratic Progressive Party, Bosnia)</td>
<td>波士尼亚 (Bosnia)</td>
<td>波士尼亚总理塔奇克的助理表示：“…”。由于...。另外，伊瓦尼奇表示，在中央政府担任要职的两名他所属的民主进步党党员也将辞职。</td>
<td>The assistant of <strong>Bosnia</strong> Premier Taqik said “…” . Because ... . In addition, <strong>Ivanic</strong> said, two <strong>Democratic Progressive Party</strong> members who held important duties in the central government…</td>
<td></td>
</tr>
</tbody>
</table>
Good Neighbors in Target KB

- Two KB entries x and y co-occur if:
  - x’s page links to y
  - y’s page links to x
  - Some KB entry z’s page links to both x and y
  - Some combination of the above
Iterative Alignment of Cross-lingual Information Networks

- It is likely that a similarity metric will assign a high score to a name mention and its KB referent.
- For some name mentions this won’t be the case.
- Named entities that co-occur in one language will co-occur in another.
- Name mention’s neighbors will often refer to correct KB entry’s neighbors.
- Low scores may be boosted based on high scoring Neighbor pairs.
Iterative Alignment of Cross-lingual Information Networks

- Follow the general approach in (You et al., 2010)
- Construct cross-lingual entity matrix based on similarity metrics
  - Edit distance between pinyin and English forms, Name transliteration, name translation
- Update initial entity pair scores iteratively:
  - $\lambda$: Interpolation parameter
  - $R$: name pair matrix
  - $t$: iteration
  - $k$: neighbor pair rank
  \[
  R_{ij}^{t+1} = \lambda \left[ \sum_{(u,v)_k \in B^t(i,j, \theta)} \frac{R_{uv}^t}{2^k} \right] + (1 - \lambda)R_{ij}^0
  \]
- Selection of translation pairs: greedy algorithm that selects pairs with highest confidence above a certain threshold and ensures one-to-one alignment
Hypothesis 2 Elaboration

Li Na

player
tennis
Russia
single
final
gain
half
female

Li Na

Pakistan
relation
express
vice president
country
Prime minister
**Fundamental Theory**: InforNet construction and knowledge discovery capability can be mutually enhanced by network analysis on text and interconnected data.

😊 **Q1**: How to discover latent topics and identify clusters of multi-typed objects simultaneously?

**A1**: Probabilistic Topic Modeling with Biased Propagation to take advantage of inter-connectivity in InforNets.
Majority voting among the queries which have the same name spelling and belong to the same topic cluster, to ensure them to link to the same KB entry.
## Evaluation Results

<table>
<thead>
<tr>
<th>Entity Type</th>
<th>Query Language</th>
<th>System</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>PER</td>
<td>English</td>
<td>Baseline</td>
<td>74.7</td>
<td>73.3</td>
<td>74.0</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>Enhanced</td>
<td><strong>76.3</strong></td>
<td><strong>76.1</strong></td>
<td><strong>76.2</strong></td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>Baseline</td>
<td>37.5</td>
<td>42.0</td>
<td>39.6</td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>Enhanced</td>
<td><strong>65.1</strong></td>
<td><strong>73.1</strong></td>
<td><strong>68.9</strong></td>
</tr>
<tr>
<td>GPE</td>
<td>English</td>
<td>Baseline</td>
<td>82.1</td>
<td>81.2</td>
<td>81.6</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>Enhanced</td>
<td><strong>82.1</strong></td>
<td><strong>82.3</strong></td>
<td><strong>82.2</strong></td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>Baseline</td>
<td>73.5</td>
<td>74.9</td>
<td>74.2</td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>Enhanced</td>
<td><strong>83.3</strong></td>
<td><strong>83.9</strong></td>
<td><strong>83.6</strong></td>
</tr>
<tr>
<td>ORG</td>
<td>English</td>
<td>Baseline</td>
<td>77.5</td>
<td>81.0</td>
<td>79.2</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>Enhanced</td>
<td><strong>80.3</strong></td>
<td><strong>84.9</strong></td>
<td><strong>82.5</strong></td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>Baseline</td>
<td>68.3</td>
<td>83.9</td>
<td>75.3</td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>Enhanced</td>
<td><strong>69.7</strong></td>
<td><strong>85.7</strong></td>
<td><strong>76.8</strong></td>
</tr>
<tr>
<td>ALL</td>
<td>English</td>
<td>Baseline</td>
<td>78.4</td>
<td>79.0</td>
<td>78.7</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>Enhanced</td>
<td><strong>79.9</strong></td>
<td><strong>81.7</strong></td>
<td><strong>80.8</strong></td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>Baseline</td>
<td>56.3</td>
<td>63.4</td>
<td>59.6</td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>Enhanced</td>
<td><strong>71.0</strong></td>
<td><strong>79.8</strong></td>
<td><strong>75.1</strong></td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>Baseline</td>
<td>63.3</td>
<td>67.6</td>
<td>65.4</td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>Enhanced</td>
<td><strong>73.6</strong></td>
<td><strong>79.9</strong></td>
<td><strong>76.6</strong></td>
</tr>
</tbody>
</table>
CLEL Knowledge Categorization

- "何伯" (He Uncle) refers to "an 81-years old man" or "He Yingjie" (5.9%)
- "丰华中文学校 (Fenghua Chinese School)" (7.6%)
- "莱赫. 卡钦斯基 (Lech Aleksander Kaczynsk) vs. 雅罗斯瓦夫. 卡钦斯基 (Jaroslaw Aleksander Kaczynski)" (4.5%)
- News reporter "Xiaoping Zhang", Ancient people "Bao Zheng" (1.8%)
- "信息网络 (Information networks)" (1.7%)
- "文档级语境 (Document-level context)" (12%)
- "话语理解 (Discourse reasoning)" (62%)
- "背景知识 (Background knowledge)"
- "无线索实体 (No-clue entities)"
- "NIL singletons"
- "流行主导实体 (Popularity-dominant entities)"
- "名字拼写 (Name spelling)"
- "表面语境 (Surface context)"
- "实体类型 (Entity type)"
Person Name Translation Challenges

Name Transliteration + Global Validation:

- 34% - 克劳斯 (Klaus), 莫科 (Moco)
  比兹利 (Beazley), 皮耶 (Pierre)…

Name Pair Mining and Matching (common foreign names)

- 28% - 伊莎贝拉 (Isabella), 斯诺 (Snow),
  林肯 (Lincoln), 亚当斯 (Adams)…

Pronunciation vs. Meaning confusion

- 3% - 拉索 (Lasso vs. Cable)
  何伯 (He Uncle)

Entity type confusion

- 3% - 魏玛 (Weimar vs. Weima)

Chinese Name vs. Foreign Name confusion

- 1.5% - 洪森 (Hun Sen vs. Hussein)

Origin confusion

- 1.5% - 王菲 (Faye Wong)

Mixture of Chinese Name vs. English Name

- 1.5% - 威恩 (Wen vs. William)

Chinese Names (Pinyin)

- 27% - 王其江 (Wang Qijiang), 吴鹏 (Wu Peng), …
Cross-lingual NIL Clustering

• One-to-Many Clustering
  • Li Na, Wallace, ...

• Topic Modeling Errors
  • The same name (莫里西/Molish), the same topic (life length/death analysis), different entities

• Require temporal employment tracking
  • 众议院情报委员会主席高斯 (Gauss, the chairman of the Intelligence Committee) = 美国中央情报局局长高斯 (The U.S. CIA director Gauss)
Morphing in Texts: KB unknown

To girlfriend:

"The first semester commences in three weeks. Two high schools and two universities... This summer will surely be hot... 19 certificates for private education and four exams. Goodbye.

boyfriend

Ramzi bin al-Shibh

on September 11

World Trade Center, Pentagon and Capitol

19 hijackers, four planes

19 certificates for private education and four exams. Goodbye.

Bin Laden

Abu Abdul Rahman
# Morphs in Social Media

<table>
<thead>
<tr>
<th>Morph</th>
<th>Target</th>
<th>Motivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blind Man</td>
<td>Chen Guangcheng</td>
<td>Sensitive</td>
</tr>
<tr>
<td>First Emperor</td>
<td>Mao Zedong</td>
<td>Vivid</td>
</tr>
<tr>
<td>Kimchi Country</td>
<td>Korea</td>
<td>Vivid</td>
</tr>
<tr>
<td>Rice Country</td>
<td>United States</td>
<td>Pronunciation</td>
</tr>
<tr>
<td>Kim Fat</td>
<td>Kim Jong-il</td>
<td>Negative</td>
</tr>
<tr>
<td>Miracle Brother</td>
<td>Wang Yongping</td>
<td>Irony</td>
</tr>
</tbody>
</table>

“Conquer West King” (平西王) = “Bo Xilai” (薄熙来) = “Baby” (宝宝) = “Wen Jiabao” (温家宝)
Morph Decoding

- Goal: automatically determine which term is used as a morph, and resolve it to its true target

- Then Wu Sangui helped the army of Qing dynasty invaded China, and became Conquer West King.

- Conquer West King from Chongqing fell from power, still need to sing red songs?
Morph Discovery

• Morph Mention Extraction
  o Identify a set of mentions which are likely to be a morph

• Morph Sense Disambiguation
  o Detect whether a morph mention is used as a morph or not in a specific message

• Why two separate steps
  o Collective disambiguation needs cleaner morph mention candidates
Mention Extraction: Features

• Basic Features
  • surface form, pos-tag, # of characters, pin-yin related features, whether characters in the term are identical
  • mostly with part-of-speech tag NN

• Dictionary-Based Features
  • many potential morphs are non-regular names, which are derived from their regular names or are newly created terms.
  • 吃省 (Eating Province) is derived from 广东省 (Guangdong Province)
Anomaly Analysis based Features

- Morphs tend to appear in anomaly contexts

**Regular Context**

- Language Modeling (LM)
  - Word-based and character-based LM
  - Use Gigaword to train the regular LM

- Context Diversity Analysis
  - Context similarity between social media and regular context
Collective Morph Sense Disambiguation

- Coreference
  - Social relations and context similarity
- Mention Correlation
  - Co-occurrence
- Semi-supervised graph regularization for collective disambiguation
Experimental Setting

• Data
  o A sample of 5363 Weibo messages
  o 6226 annotated morph mentions, and 450 unique mentions, 5210 out of 6226 are morphs

• Evaluation Metrics
  o standard precision, recall and F1
Preliminary Results (Huang et al, EMNLP2014 submission)

• Morph Mention Extraction

<table>
<thead>
<tr>
<th>Features</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>0.302</td>
<td>0.709</td>
<td>0.424</td>
</tr>
<tr>
<td>+Dictionary</td>
<td>0.657</td>
<td>0.660</td>
<td>0.658</td>
</tr>
<tr>
<td>+Diversity</td>
<td>0.743</td>
<td>0.649</td>
<td>0.693</td>
</tr>
</tbody>
</table>

• Morph Sense Disambiguation
  - Naïve: consider all mentions as morphs
  - SSRegu: our proposed collective inference model

<table>
<thead>
<tr>
<th>Methods</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>0.76</td>
<td>0.87</td>
<td>0.812</td>
</tr>
<tr>
<td>SSRegu</td>
<td>0.883</td>
<td>0.862</td>
<td>0.872</td>
</tr>
</tbody>
</table>
Target Candidate Identification

- Considering all entities will be too overwhelming
  - Make resolution difficult and affect system efficiency
- Temporal Distribution Assumption
  - Intuition: social users should know the real targets before they use morphs
  - Assume the target candidates should appear within certain time period (e.g., 7 days) of the morph
  - Naïve but greatly narrow down candidates into 1% and keep 92% of all targets
Target Candidate Ranking: Motivating Example

- **Conquer West King** from Chongqing fell from power, still need to *sing red songs*?
- There is no difference between that guy’s plagiarism and **Buhou**’s *gang crackdown*.
- Remember that **Buhou** said that his family was not rich at the press conference a few days before he fell from power. His son **Bo Guagua** is supported by his scholarship.

- **Bo Xilai**: ten thousand letters of accusation have been received during Chongqing gang crackdown.
- The webpage of “Tianze Economic Study Institute” owned by the liberal party has been closed. This is the first affected website of the liberal party after **Bo Xilai** fell from power.
- **Bo Xilai** gave an explanation about the source of his son, **Bo Guagua**’s tuition.
- **Bo Xilai** led Chongqing city leaders and 40 district and county party and government leaders to *sing red songs*.

MORE

Weibo (censored)

Twitter and Chinese News (uncensored)
Heterogeneous Information Network

Example of Morph-Related Heterogeneous Information Network

Network Schema

M: Morphs
E: Entities
EV: Events
NP: Non-Entity Noun Phrases

- Three types of Meta-paths:
  - M – E – E
  - M – EV – E
  - M – NP – E

- Each meta-path provides a unique angle to measure how similar two objects are
Meta Path-based Similarity Measures

- Common Neighbors: the number of common neighbors between a morph \( m \) and a target \( e \)
- Path Count: the number of paths between \( m \) and \( e \)
- Pairwise Random Walk
  \[ \text{sim}(m, e) = \sum_{(p_1p_2) \in (P_1P_2)} \text{prob}(p_1)\text{prob}(p_2^{-1}) \]
  - Favor nodes with highly skewed distribution in in-links/out-links
- Kullback-Leibler Distance
  \[ \text{sim}(m, e) = \sum_{i=1}^{N} p_m(x_i) \log \frac{p_m(x_i)}{p_e(x_i)} + p_e(x_i) \log \frac{p_e(x_i)}{p_m(x_i)} \]
Two New Similarity Measures

- Based on Cosine similarity-style normalization method
- Normalized common neighbors

\[ sim(m, t) = \frac{\left| \Gamma(m) \cap \Gamma(t) \right|}{\sqrt{\left| \Gamma(m) \right|} \sqrt{\left| \Gamma(t) \right|}} \]

- refines the simple counting of common neighbors by avoiding bias to highly visible or concentrated objects
- Pairwise random walk/cosine

\[ sim(m, t) = \sum_{p1p2 \in (P1 P2)} f(p1)f(p2^{-1}) \]

- Where \( f(p1) = \frac{\text{count}(m, x)}{\sqrt{\sum_{x \in \Omega} \text{count}(m, x)}} \)

\[ f(p2) = \frac{\text{count}(t, x)}{\sqrt{\sum_{x \in \Omega} \text{count}(t, x)}} \]

- A tamer version of pairwise random walk
Global Semantic Features

- A morph tends to have higher temporal correlation with its real target, and share more similar topics compared to other irrelevant targets.
- Capture such two dimension information simultaneously may be helpful.
- Two types of global features:

\[ \text{sim}_{\text{global sum}}(m, e) = \sum_{t_i \in T} \text{sim}_{t_i}(m, e) \]

\[ \text{sim}_{\text{global norm}}(m, e) = \sum_{t_i \in T} \text{norm}_{t_i}(m, e). \]

- Where

\[ \text{norm}_{t_i}(m, e) = \frac{\text{sim}_{t_i}(m, e)}{\sum_{e \in E} \text{sim}_{t_i}(m, e)}. \]

\[ T = t_1 \cup t_2 \cup \ldots \cup t_{N_l} \text{ set of temporal slots} \]
Integrate Cross Source/Cross Genre Information

• Comparisons of Weibo and Twitter
  • Weibo: Already put in prison, still need to serve Buhou?
  • Twitter: ...call Bo Xilai “conquer west king” or “buhou”...
  • Information from media not under censorships is more explicit
  • Integrate information from Twitter to help morph resolution

• Integrate information from cross genre web documents
  • Richer and cleaner information
  • Existing NLP tools work better
Social Network Features

- Morphs tend to be spread within a community
- A morph and its target are likely to be posted by two users with strong social correlation
- Explicit Social Correlation between users from re-tweeting and mentioning networks (Wen and Lin, 2010)
  - compute the degree of separation in user interactions and the amount of interactions
  - Whether both the morph and its candidate appear in the posts of two users with strong social correlation as additional feature
Learning-to-Rank

- Logistic Regression model to combine different set of features

<table>
<thead>
<tr>
<th>Morph</th>
<th>Target</th>
<th>LCS</th>
<th>CN</th>
<th>PRW</th>
<th>Social</th>
<th>...</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conquer West King</td>
<td>Bo Xilai</td>
<td>0</td>
<td>100</td>
<td>0.4</td>
<td>0.6</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>Conquer West King</td>
<td>Wang Lijun</td>
<td>1</td>
<td>50</td>
<td>0.3</td>
<td>0.6</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>Conquer West King</td>
<td>Obama</td>
<td>0</td>
<td>4</td>
<td>0.001</td>
<td>0.0</td>
<td>...</td>
<td>0</td>
</tr>
</tbody>
</table>
Data and Scoring Metric

- **Data**
  - Time frame: 05/01/2012-06/30/2012
  - 1555K Chinese messages from Weibo
  - 66K formal web documents from embedded URL
  - 25K Chinese messages from English Twitter for sensitive morphs
  - Test on 107 morph entities in Weibo, 23 of them are sensitive

- **Scoring Metric**
  \[ Acc@k = \frac{C_k}{T} \]
  - \( C_k \): the number of correctly resolved morphs at top position \( K \)
  - \( T \): the total number of morphs in ground truth
Overall performance (Huang et al, ACL2013)
Effects of Features

- 1: Similarity over homogeneous networks (Hsiung et al. 2005)
- 2: Similarity over Heterogeneous Networks
- 3: 2 + two new similarity measures
- 4: 3 + global features
- 5: 4 + cross genre/sources information
- 6: 5 + social features