TAC Entity Linking by Performing Full-document Entity Extraction and Disambiguation

Silviu Cucerzan
Microsoft Research Machine Learning Group

Gaithersburg, MD
KBP Entity Linking - Task Description

For a name string and a document, determine which entity in a given knowledge base if any is being referred to by the name string.

Evaluation metrics:
Linking accuracy ($A$), Known-entity linking accuracy ($A_{Wiki}$), NIL accuracy ($A_{NIL}$), B-cubed precision and recall with equal element weighting

$$P_{B\text{-cubed}^+} = \text{Avg}_x(\text{Avg}_{x'}|T(x)=T(x')\delta(T(x),S(x),S(x')))$$

$$R_{B\text{-cubed}^+} = \text{Avg}_x(\text{Avg}_{x'}|S(x)=S(x')\delta(S(x),T(x),T(x')))$$


... → E0421510: Reserve Bank of Australia

... → E0700143: Reserve Bank of India

... → NIL
Employed Resources


Text corpus:

Annotated data:

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Size (entity mentions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>Organization</td>
</tr>
<tr>
<td>2010 Training Web</td>
<td>500</td>
</tr>
<tr>
<td>2010 Eval Newswire</td>
<td>500</td>
</tr>
<tr>
<td>2010 Eval Web data</td>
<td>250</td>
</tr>
</tbody>
</table>

1 million news articles
+ 300,000 Web documents

818,741 nodes

3.6 million nodes
How Ambiguous Are Target Names?


- E0421510: Reserve Bank of Australia
- E0700143: Reserve Bank of India

NIL
NEW DELHI/MUMBAI: Layoffs, firings and salary cuts are increasingly becoming all too common across India Inc, highlighting a deepening slowdown in the economy that has forced companies to take the knife to costs to protect their bottom line.

From banking and finance to aviation, from manufacturing to information technology, no sector appears immune, as companies look beyond hiring freezes to job cuts, mirroring a trend across much of the developed world which has seen tens of thousands of people out of employment.

Admittedly India, among the few major global economies that will see respectable GDP growth this year, may not see job losses quite like that being felt in the West, it has nevertheless got policymakers worried.

Prime Minister Manmohan Singh earlier this week urged industry to desist from laying off people and promised to cut interest rates and levies to shore up the economy. The Reserve Bank of India (RBI) has already turned its attention to driving up growth from containing inflation, and cut key reserve ratios for banks and a short-term interest rate, signalling a bias in favour of lower rates.

Yet on Friday, news about job cuts came in from different directions. L&T Infotech, a wholly-owned subsidiary of the country’s largest engineering company Larsen & Toubro (L&T), is shedding up to 5% of its workforce of nearly 10,000 employees, according to market sources.

[...]
How Ambiguous Are Target Names?

<query id="EL006455">
  <name>Reserve Bank</name>
  <docid>eng-NG-31-100316-11150589</docid>
  <entity>E0700143</entity>
</query>

<query id="EL06472">
  <name>Reserve Bank</name>
  <docid>eng-NG-31-142262-10040510</docid>
  <entity>E0421510</entity>
</query>

Wiki Oct. 2008

... → E0421510: Reserve Bank of Australia
... → E0700143: Reserve Bank of India
... → NIL

Wiki Oct. 2008: “reserve bank”
8 entities

Wiki June 2011: “reserve bank”
9 entities
105 surface forms that contain the string “reserve bank”
68 entities
Full-document Analysis

- Perform full-document entity extraction and then match target names against the extracted entities; choose the top-ranked matching entity.
- Sub- and super-string matches (in 7% of the instances, the target name does not match exactly any entity reference extracted from the text).

  e.g. “USC” ~ “USC baseball team” \(\rightarrow\) USC Trojans baseball

  “Koran Tempo newspaper” ~ “Koran Tempo” \(\rightarrow\) Koran Tempo
Starting Point

- Productized concept extraction system, trained on Wikipedia from June 2011

- Map the provided KB extracted from Oct 2008 data to the 2011 collection

2010 training data set: A = 86.3%

2010 training data set: A = 88.2%
Overview of the Information Extracted

Surface Forms

- Texas
- Texas (US State)
- University of Texas Austin
- USS Texas
- Texas (band)
- Texas (musical)
- Texas (TV Series)
- Texas (novel)
- Texas (SpongeBob episode)
- Texas Instruments
- Texas County, OK

≈ 30

Entities

- Texas (TV Series)

Topics:
- NBC network shows
- American television soaps
- Television spin-offs

Contexts:
- Another World
- TV Series

e.g.: Texas
Surface Form to Entity Mappings

• the titles of entity pages
Surface Form to Entity Mappings

- the titles of redirecting pages

http://en.wikipedia.org/wiki/Another_World_in_Texas
Surface Form to Entity Mappings

- the disambiguation pages
Surface Form to Entity Mappings

- the references to entity pages in other articles

Pam Long

From Wikipedia, the free encyclopedia

Pamela K. Long is an American actress and writer, most known for her stints writing the soap opera Guiding Light, from 1983 to 1986 and from 1987 to 1990.

Before she started in show business, she won the title of Miss Alabama in 1974. At the time she was attending the University of North Alabama, and was involved in the Phi Mu sorority.

After graduation, she went to New York City and played Ashley Linden on Texas from 1981 to 1982. It was after the show ended that Long, who had already proven herself capable of writing a soap, was penning scripts for Texas, moved to Guiding Light with former Texas executive producer Gail Kobe.

Both of her stints were well-regarded and provided a renaissance for the once-sliding show, at one point pushing the show's ratings to the #1 spot for three weeks in the summer of 1984, dethroning then-powerhouse General Hospital from the top ratings spot. Long created some of GL's most memorable characters, like vixen Reva Shayne, haughty ice queen Alexandra Spaulding, and rough-around-the-edges ingenue Harley Davidson Cooper.

When she left Guiding Light for the second time, she went on to write Santa Barbara and One Life to Live.

Long was once married to actor Jay Hammer and was credited for a time as Pamela Long Hammer.

<table>
<thead>
<tr>
<th>Preceded by</th>
<th>Head Writer of Guiding Light (GL)</th>
<th>Succeeded by</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carolyn Culliton</td>
<td>1983-1986</td>
<td>Jeff Ryder</td>
</tr>
<tr>
<td>Preceded by</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Joseph D. Manetta</td>
<td>Head Writer of GL</td>
<td>Nancy Curlee</td>
</tr>
<tr>
<td></td>
<td>1987-1990</td>
<td>Stephen Demorest</td>
</tr>
</tbody>
</table>
Topics

- List pages ("List of [...]" "Table of [...]")
Texas (TV series)

From Wikipedia, the free encyclopedia

Texas (sometimes called Another World in Texas and, later, Texas: The New Generation) was a soap opera which aired on NBC from August 4, 1980 until December 31, 1982. Created by John William Corrington, Joyce Hooper Corrington, and Paul Rauch, the show was a spinoff of Another World. It was unique in that it was the first soap opera to air hour-long episodes from its inception. The show aired from 3pm-4pm (EST) and caused a small domino effect on the NBC daytime schedule; Another World, which had previously been airing from 2:30pm-4pm, was scaled back to 60 minutes and aired from 2pm-3pm. The Doctors, which had been airing from 2pm-2:30pm, moved to 12:30pm-1pm.

Texas was moved towards the end of its run, taking the 11 AM to noon slot on the schedule. This lasted from April to December of 1982, when Texas was cancelled.

Rauch’s initial plans involved a historical soap centered around the period of the Civil War. The Corringtons submitted a concept for a show set in the Antebellum South entitled “Reunion”, but NBC wanted something more in line with the hugely successful primetime soap Dallas. Rauch then chose to have the show revolve around the popular

The second was the Match Game-Hollywood Squares Hour, hosted by Gene Rayburn and Jon Bauman. A soap opera wouldn’t air in that slot again until July 1984, when Santa Barbara premiered. After that was cancelled in 1993, NBC gave the hour back to its affiliates.

Categories: 1960s American television series | NBC network shows | American television soap operas | Television spin-offs | Television series by Procter & Gamble Productions | Television shows set in Texas
Topics

• Lexico-syntactic patterns
• List pages: 80k
• Categories: 456k
• Lexico-syntactic patterns: 852k

• Avg. # topics per entity: 4.5
• Avg. # entities per topic: 12
Disambiguation - Intuition

Let $C = \{c_1, ..., c_M\}$ be the set of known contexts, and $T = \{t_1, ..., t_N\}$ be the set of known topic identifiers.

Given a text document $D$, our goal is to maximize the similarity between the document context $d$ and each entity's contexts as well as the topic identifiers of each entity pair.

S. Cucerzan. “Large-scale Entity Disambiguation Based on Wikipedia Data”. EMNLP 2007

Maximize the similarity between the document context $d$ and each entity's contexts as well as the topic identifiers of each entity pair.
Disambiguation - Intuition

S. Cucerzan. “Large-scale Entity Disambiguation Based on Wikipedia Data”. EMNLP 2007

$$\arg \max \quad \beta \sum_{i=1}^{n} < C_{e_i}, d > + (1 - \beta) \sum_{i=1}^{n} \sum_{j=1 \neq i}^{n} < T_{e_i}, T_{e_j} >$$

$$\overline{d} = d \times \bigcup_{s \in S(D)} \bigcup_{e \in \varepsilon(s)} T_{e}$$

$$\arg \max \quad \sum_{i=1}^{n} < \left( \frac{\beta}{1 - \beta} C_{e_i}, T_{e_i} \right), \overline{d} - (0, T_{e_i}) >$$

$$\arg \max \quad < \left( \frac{\beta}{1 - \beta} C_{e_i}, T_{e_i} \right), \overline{d} > - \| T_{e_i} \|^2, i \in 1..n$$

# topic tags of $e_i$
I guess all that praying and rehab is helping out Maia Campbell! Here she is in a new independent film alongside the likes of Ray J, LisaRaye, AZ and more. The movie is called "Envy" and will be out on DVD on November 10th. All I'm going to say to Ray J is......keep your day job at VH1! Check out the movie trailer below:
Disambiguation Component

Text document $D$

$D \cap \bigcup_{s \in \varepsilon(s)} C(e), D \cap \bigcup_{s \in \varepsilon(s)} V_T(e), \bigcup_{s \in \varepsilon(s)} T(e), \bigcup_{s \in \varepsilon(s)} V_T(e)$

$\{w | w \in T(e)\}$

$S_1$ → $e_1^{s_1}, \ldots, e_{|\varepsilon(s_1)|}^{s_1}$

$S_i$ → $e_1^{s_i}, \ldots, e_k^{s_i}, \ldots, e_{|\varepsilon(s_i)|}^{s_i}$

$C(e_k^{s_i}), T(e_k^{s_i}), V_T(e_k^{s_i})$

$S_j$ → $e_1^{s_j}, \ldots, e_l^{s_j}, \ldots, e_{|\varepsilon(s_j)|}^{s_j}$

$C(e_k^{s_j}), T(e_k^{s_j}), V_T(e_k^{s_j})$

$S_n$ → $e_1^{s_n}, \ldots, e_{|\varepsilon(s_n)|}^{s_n}$

$\{w | w \in T(e)\}$
Disambiguation Component

\[
\arg \max_{e_i \in \varepsilon(s_i)} \sum_{j=1..|F|} \beta_j \cdot f_j(e_i, \bar{D}), i \in 1..n
\]

Features:
- Wikipedia-based prior
- Context similarity with the document
- Topic-identifier similarity
- Topic-word similarity with the document
- Topic-word similarity
- Number of different surface forms in the document that lead to the same entity
- Acronym matching *
- String similarity
- Required context not found in the document
- Unlikely type of entity (music, movies, etc.) *
Training: Binary Labeled Examples

• Entity in the 2008 collection:
  – Label 1 for the correct disambiguation
  – Label 0 for all other possible disambiguations

• NIL entity
  – Label 0 for all possible disambiguations in the 2008 collection
Target Names vs. Surface Forms

• Perform full-document entity extraction and disambiguation, which includes coreference and partial name resolution
• Map the target string to one of the extracted entities by matching it against the extracted surface forms for the entities.
Coreference and Partial Name Resolution

Coreference → positional and string matching heuristics to map shorter surface forms to longer surface forms labeled with the same entity type

Partial names → use the complementary parts of the entities of the type person identified in the document to create disambiguation candidates
The jury in the Michael Jackson molestation trial deliberated for about 2 hours yesterday afternoon before recessing for the day... they could come back with a verdict at any time now. But the really important thing to discuss is way that Janet Jackson looked when she showed up at court to support her brother yesterday:

Um, Janet has been packin' on the lbs. She still looks good but you can tell that she is getting big. Is it wrong to say that she is starting to look like one of the Klumps?

Everyone is talking about how amazing Jessica Simpson looks in her new video These Boots are Made for Walking... and while I will not deny that she looks absolutely killer I do feel it is my duty to point out one tiny little thing:

Er... what is going on in there? Do we even want to know?

Yay! They are coming out with a David Beckham toy:

It looks fun to play with... but I think I might prefer the real thing:

Oh c'mon, you know you would love to play with his ball too.

Who knew Andy Dick had a son?

Yeah, I feel bad for the kid... can you imagine having a Dick for a dad?

Ashanti... hahahaha... I don't even know what to say...

... she needs to take that wig back and get her $20 back... it's SOOOO horrible!!!
<query id="EL04014">
  <name>Dick</name>
  <docid>eng-WL-11-174595-12967356</docid>
  <entity>E0039314</entity>
</query>

<query id="EL04015">
  <name>Dick</name>
  <docid>eng-WL-11-174595-12967728</docid>
  <entity>E0111000</entity>
</query>

<query id="EL04018">
  <name>Dick</name>
  <docid>eng-WL-11-174601-12969743</docid>
  <entity>NIL</entity>
</query>

<query id="EL04019">
  <name>Dick</name>
  <docid>eng-WL-11-174643-13000483</docid>
  <entity>NIL</entity>
</query>

<DOCID>eng-WL-11-174595-12967728</DOCID>
<DOCTYPE SOURCE="blog">BLOG TEXT</DOCTYPE>
<DATETIME>2006-03-24T18:01:00</DATETIME>
<HEADLINE>Turn That Frown Upside Down</HEADLINE>
<TEXT>
  <POST>
    <POSTER>???</POSTER>
    <POSTDATE>2006-03-24T18:01:00</POSTDATE>
    Hmm ... I wonder what is going on with our dear Jake Gyllenhaal ... he's been lookin' mighty sad the past few days he's been out and about. Check out these pictures of Jake y poo looking all sad and stuff ... I wonder what it will take to get a smile on his face? [...] Parker Posey and Chris Kattan were on hand in NYC last night for the premiere of their new movie Adam &amp; Steve:
    I love Parker so much! This movie doesn't sound that amazing but she is so damn funny I'll prolly check it out just because she's in it. I'm really excited to see her in Superman Returns this June. [ Source , Source ]
    UGH ... Sharon Stone really needs to chill out with the over-the-topness she's been attempting lately ... check out these ridiculous pictures from the German premiere of Basic Instinct 2:
    [...] Rumors have been swirling lately that Paris Hilton and Stavros Niarchos were dunzo ... that she couldn't handle his Greekness and went back to her last boyfriend Paris Latsis (who is also Greek ... yeah ... I know ). But check out this picture of Parisopolis looking all cozy while out shopping in LA:
    [...] TheSmokingGun.com has sniffed out the official tour rider for Vice President Cheney ... now we know what sorts of things the VP needs in his hotel suites to make him happy: As far as tour riders, this one isn't too bad (I think J. Lo has the most ridiculous rider I've ever seen altho, Mariah Carey 's rider is pretty insane too). I love that all the TVs have to be tuned to Fox News ... so typical. Do you suppose Dick "accidentally" shoots people who forget to give him his 4 cans of Diet Caffeine Free Sprite? [ Source ]
  </POST>
</TEXT>
Netflix Review! THIS FILM IS NOT YET RATED (2006, dir. Kirby Dick)

To quote another more successful documentary from 2006, "It's not so much a political issue as it is a moral issue..." The subject of this doc is the mysterious and fear-inspiring MPAA rating system. I got a call today at work from a producer who uttered the phrase "I just heard from the MPAA" with a pathetic quiver reminiscent of a child waking from the nightmare. Filmmaker Kirby Dick gives the hard-hitting expose structure a lighthearted touch that works to a degree but ultimately keeps any non-cinephile at a comfortable distance.

What is the MPAA? Who makes up the rating board? The appeals board? These questions and more are answered by the amiable Dick and a crafty P.I. named Becky who looks like the best 3rd grade gym teacher ever! It actually is a very CRUEL injustice that practically every great filmmaker of the last 50 years has had to recut their film to avoid an "X" or "NC-17" in order to secure distribution in the United States. What the film doesn't answer and actually can't answer is exactly what criteria constitutes this rating. The whole rating process is shrouded in secrecy. The film does make some deductions that any cinephile probably already knows.

OK stuff: shooting people in the head. stabbing a woman in her fake breast. rapid gun fire with no blood. rape. NOT OK stuff: female orgasms. pubic hair. pouring spagetti all over your breasts. realistic depictions of war crimes. So... what's the result? Well, you get movies that promote video-game violence and discourage consensual sex. It's backwards. It's weird. We all agree. What the documentary strains to also point out is that it is indeed censorship. It's a bit of a violation of that ye olde first amendment. But the larger issue is that film is not considered art. It should be but it's not. It's a business. [...]
She remembers her son Teddy Ebersol.

The actress and wife of NBC Sports chairman Dick talks about the 14-year-old's tragic death.

"Teddy was funny. The youngest one—we were looking at environmental factors and (aircraft) performance factors," Arnold Scott said.

Dick Ebersol of sports Meets of NBS has told researchers, his diplomaed jet plane struggled only 20 ft in air before it has fallen back to a runway and has broken separately, officials of aircraft have told on Wednesday.

14-years son Edward Ebersol "Teddy" and two members of a team was killed, when the engine of twin CL-601 the Applicant has failed in easy snow, a fog and freeze temperatures in Montrose the Regional Airport on Sunday.

National researchers of Board of Safety of Transportation have told, that the plane has not removed ice before attempt of rise. (from news - 2004)

"Teddy was funny. The youngest one— he's just there, you know? And we had such a big family. We have five children. Teddy was kind of like— Teddy, come on, Teddy. Hurry up, Teddy. Quietly he developed this way of thinking that you would never know about, except the school he went to makes them write an autobiography. And so he wrote this autobiography. At the time— we read it found out things we didn't know about Teddy and how much we all meant to him as a family—" one quote said something about, "On the road of life, the only person ahead of the love of your family is God."

[...]

Dick Ebersol
More than 10% target names in TAC 2010 data.

- Israeli Air Force
- Islamic Academy of Florida (non-Wiki)
- Israeli Air Force
- Indian Air Force
Syria has not yet accepted a request by the International Atomic Energy Agency to visit the site bombed by the IAF on September 6, which Washington says was a nuclear reactor, Reuters reported Friday.

The news agency quoted diplomats in Vienna as saying that Damascus was stalling its approval of the UN delegation visit, demanding more details on the proposed inspection.

Syrian atomic energy chief Ibrahim Othman came to Vienna earlier this month to speak with IAEA head Mohamed ElBaradei on the matter, but the two did not agree on the timing or nature of a visit, diplomats said.

The agency received a letter from Syria several days ago asking for more details on the trip, one diplomat said. The IAEA has replied and is now waiting for Damascus's response, he added.
This weekend, the University of South Florida (USF) and the town of Temple Terrace, a suburb of Tampa Bay, Florida, will once again be hosting the Muslim American Society's Olympics. Once again, the school known as Jihad U and the local area will be associating themselves with terrorism and radical Islam.

MAS-Tampa website has been purged of its hate and violence, USF didn't have a problem hosting the radical group when the material was up on the site in 2005 and 2006. But then again, radical Islam has become a normal occurrence at USF.

Sami Al-Arian arrived at USF in 1986. While there, he would create an entire infrastructure for Palestinian Islamic Jihad (PIJ), an organization that targets Israeli civilians with terrorist attacks. This PIJ network consisted of a charity, the Islamic Committee for Palestine (ICP) a.k.a. Islamic Concern Project; a think tank, World and Islam Studies Enterprise (WISE); and a children's school, the Islamic Academy of Florida (IAF). In the case of WISE, USF was a partner.

Ramadan Shallah, at the behest of Al-Arian, came to teach a political science course at USF, in 1991. He abruptly left town in 1995 and soon emerged in Damascus, Syria as the international head of PIJ, replacing the assassinated Fathi Shikaki.

[...]

Islamic Academy of Florida (IAF)
IAEA finds enriched uranium in Syria....

According to Monday's report, the uranium was discovered at the same site which was allegedly bombed by IAF jets in September 2007. Behind the scenes, Israel has reportedly been working to convince US and other Western officials of the legitimacy of the air strike, but the findings of the IAEA investigators provide the first independent confirmation that a nuclear program had indeed been in development.

The leaked information came shortly after the IAEA Director Mohamed ElBaradei announced he would release a formal, written report on the subject, Reuters reported. The IAEA had no immediate comment.
Take care of your parents!!

THEY ARE PRECIOUS

This was narrated by an IAF pilot to IIT students during a Seminar on Human Relations:

Venkatesh Balasubramaniam (who works for IIT) describes how his gesture of booking an air ticket for his father, his maiden flight, brought forth a rush of emotions and made him (Venkatesh) realize that how much we all take for granted when it comes to our parents.

My parents left for our native place on Thursday and we went to the airport to see them off. In fact, my father had never traveled by air before, so I just took this opportunity to make him experience the same. In spite of being asked to book tickets by train, I got them tickets on Jet Airways. The moment I handed over the tickets to him, he was surprised to see that I had booked them by air. The excitement was very apparent on his face, waiting for the time of travel. Just like a school boy, he was preparing himself on that day and we all went to the airport, right from using the trolley for his luggage, the baggage check-in and asking for a window seat and waiting restlessly for the security check-in to happen. He was thoroughly enjoying himself and I, too, was overcome with joy watching him experience all these things.

As they were about to go in for the security check-in, he walked up to me with tears in his eyes and thanked me. He became very emotional and it was not as if I had done something great but the fact that this meant a great deal to him.
Acronyms

IAF

Web / Search data:
israeli air force
international association of facilitators
institute for alternative futures
industrial areas foundation
international accreditation forum
inter american foundation
israel air force
integrated architecture framework
intelligent audio file
iraqi accordance front
inspired art fair
international astronautical federation
international advertising festival
indian air force
islamic action front
infrastructure assessment framework
industrial air filtration
international apparel federation
islamic academy of florida
integration adapter framework
international of anarchist federations
inject a floor
international academy of flint

Wikipedia – June 2011 extract:
IAF (disambiguation)
Israeli Air Force
Indian Air Force
Indonesian Air Force
International Accreditation Forum
International Astronautical Federation
Islamic Action Front

restrict the candidate space to only acronym expansions present in the target document

Other Subtasks

- Truecasing of titles and other capitalized content
- Concept boundary detection
- Topic segmentation / relevant context identification
- Resolving conflicting disambiguations (within a document)
- NIL prediction
- NIL clustering (only based on Wikipedia 2011 and exact surface form matching)
Submitted System Performance

- 2010 data:
  - Training
    A = 89.93%
    Ent: 90.6% NIL: 88.3%
  - Test
    A = 89.96%
    Ent: 87.3% NIL: 92.2%

- 2011 official run:
  \[ F_{B\text{-cubed}^+} = 0.841 \]
  A = 86.8%
Conclusion

- full entity analysis of a target document by modeling it in the space of Wikipedia-derived topics and contexts of all candidate entity disambiguations for all surface forms extracted from the target document
- empirical results obtained on the test set suggest that this system is achieving current state-of-the-art entity linking performance