

Stanford-UBC at TAC-KBP

Eneko Agirre, Angel Chang,
Dan Jurafsky, Christopher Manning,
Valentin Spitkovsky, Eric Yeh



Ixa NLP group, University of the Basque Country
NLP group, Stanford University



Outline

- Entity linking
- Slot filling

Outline

- Entity linking
 - Slot filling

Entity linking

string

entity

Paul Newman

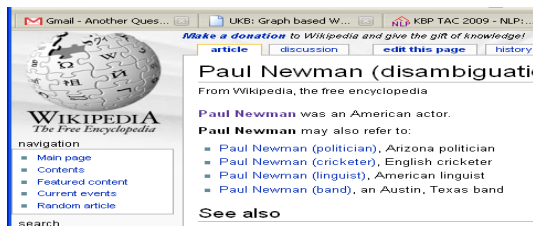
E0181364

- Given Knowledge Base
- Given target string and surrounding text:

I watched "Slapshot", the 1977 hockey classic starring Paul Newman for the first time.

- Return entity in KB (E0181364) or NIL
- KB subset of Wikipedia

Paul_Newman	E0181364
Paul_Newman_(politician)	NIL
Paul_Newman_(cricketer)	NIL
Paul_Newman_(linguist)	NIL
Paul_Newman_(band)	NIL



Entity Linking vs. Word Sense Disambiguation

- Same layout as WSD
 - Given a preexisting dictionary (sense inventory):

<i>string</i>		<i>concept</i>
counterfeit	n-03562262	monosemy
forgery	n-03562262	variants
bank	n-09213565	
bank	n-08420278	polysemy

→ Decide appropriate sense in context

He cashed a check at the bank

- Plethora of methods (Agirre and Edmonds, 2006)

Entity linking vs. Word Sense Disambiguation

- Entity linking has same layout, but...
 - Entities rather than concepts (instance vs. class)

Norfolk also took the Minor Counties One-day Title, in 1986 (under Quorn Handley) and again (at Lord's, under Paul Newman) in 1997 and 2001.
 - Dictionary is partial, needs to be completed
 - No full set of entities: those given by KB, otherwise NIL
 - Only one string, potentially many other variants
([Paul Leonard Newman](#), [Paul L. Newman](#), etc.)
- Some differences, but same techniques might work

Approaches to entity linking

- Dictionary lookup (no use of context)
 - Construct dictionary
 - Record preferred entity (prior)
- Supervised system
 - Use training examples from Wikipedia
- Knowledge-based system
 - Similarity between context and KB entry (Wikipedia article)
- Combination

Constructing the dictionary

- Table with all possible string – entity pairs
- Two purposes
 - Inventory for supervised and knowledge-base algorithms
 - Disambiguation method, using an estimation of the prior
- Space of concepts: KB concepts + Wikipedia articles
 - Remove redirection, disambiguation, list_of pages
 - Redirects are clustered, choosing KB entry as canonical form
- Space of strings: names in KB, titles of articles, plus ...
 - Redirects ([Paul Leonard Newman](#))
 - Anchor text of links to the article ([Newman](#), [Paul L. Newman](#))
 - Case normalization, fuzzy match for variations, misspellings ([Paul Newma](#))

Constructing the dictionary: priors

- For every unique string,
distribution as anchor of entity

The Prize is a 1963 spy film starring
Paul Newman ...



w) inter-Wikipedia links (03/09 dump)

W) external Web links into Wikipedia (06/09 crawl)

Paul Newman	0.9959	Paul_Newman	W:1986/1988 w:990/1000
Paul Newman	0.0023	Paul_Newman_(band)	w:7/1000
Paul Newman	0.0003	Cool_Hand_Luke	W:1/1988
Paul Newman	0.0003	Newman's_Own	W:1/1988
Paul Newman	0.0003	Paul_Newman_(austr...)	w:1/1000
Paul Newman	0.0003	Paul_Newman_(musician)	w:1/1000
Paul Newman	0.0003	Paul_Newman_(professor)	w:1/1000
Paul Newman	0	Paul_Newman_(cricketer)	
Paul Newman	0	Paul_Newman_(linguist)	
Paul Newman	0	Paul_Newman_(politician)	

Constructing the dictionary

- Three versions, depending on string matching:
 - a) EXCT: exact match
 - b) LNRM: lower-cased normalized UTF-8, minus non-alpha-numeric low ASCII
 - c) FUZZ: nearest non-zero Hamming distance matches
- Additional dictionary:
 - d) GOOG: google search site:en.wikipedia.org

Supervised disambiguation

- Given target string and surrounding text, pick most appropriate entity
 - One multi-class classifier for each target string
- Construct training data
 - Use anchors in Wikipedia text

The Prize is a 1963 spy film starring
Paul Newman ...

- Some strings have few occurrences
 - Also use other strings for the target entities
e.g for “Paul L. Newman”, also use “Paul Newman”

Supervised disambiguation

- Build multi-class classifiers for each string
 - Inspired on WSD literature
 - Features
 - Patterns around target: wordforms / lemma / PoS
 - Bag of words: lemmas in context window
 - Noun/verb/adjective before/after the anchor text
 - SVM linear kernel

Knowledge-Based disambiguation

- Given target string and surrounding text, pick most appropriate entity
 - Overlap between context and article text (Lesk, 86)
- Convert article text into a TF-IDF vector, and store into Lucene.
- Given string and text, rank articles by cosine similarity values.
 - Keep only articles in EXACT dictionary.
 - Document context:
gather 25 tokens around all occurrences of target string

Combination

- Each method outputs entities with scores
- Heuristic combinations
 - RUN1 – Cascade of dictionaries:
exact lookup, if not lower case norm, if not fuzzy
 - RUN2 – Vote using inverse of ranking
 - Cascade of dictionary
 - Google ranking
 - Supervised system
 - Knowledge-based system
- Meta-classifier
 - RUN3: Linear combination, optimized on development set using conjugate gradient

Results & Conclusions


	micro	KB
Best	82.17	77.25
Stanford_UBC2 (voting)	78.84	75.88
Stanford_UBC3 (meta)	75.10	73.25
Stanford_UBC1 (dict)	74.85	69.49
Median	71.80	63.52

- Good results overall
 - _Dictionary as cornerstone
 - _Priors remarkable
 - _NIL too conservative
- Combination
 - _Effective use of context
 - _Voting worked best
 - _Meta-classifier weak
- WSD techniques work
- Currently
 - _Error analysis

Slot filling

- Distant supervision (Mintz et al. 09):
 - Use facts in Knowledge Base (via provided mapping)
=> **gold-standard entity–slot–filler**
 - Search for spans containing entity–filler pair in document base
=> **positive examples to train**
 - Search for mentions of target entity in document collection
 - Run each of the classifiers
- Manual work kept to a minimum: types of fillers

Paul Newman



in 2007

Born	Paul Leonard Newman January 26, 1925 Shaker Heights, Ohio, U.S.
Died	September 26, 2008 (aged 83) Westport, Connecticut, U.S.
Occupation	Actor, director, humanitarian, entrepreneur
Years active	1952–2007
Spouse(s)	Jackie Witte (1949–1958) (divorced) Joanne Woodward (1958–2008) (his death)

Get gold tuples from KB

- Infobox slot names
 - Use mapping provided by organizers
 - Paul_Newman – **occupation** – “actor”
 - Paul_Newman – **per:title** – “actor”
- Ambiguity in mapping, multiple fillers in string
 - per:place_of_birth**
 - “**November 29, 1970** (1970-11-29) (age 38) Las Vegas, Nevada”
 - Set type of filler (or closed list) for each slot
 - Use NER on filler text

Train classifiers for each slot

- Extract positive examples from document base
 - 5words **entity** 0–10words **filler** 5words
 - 5words **filler** 0–10words **entity** 5words
- Negative example
 - Spans from other slots matching the entity type (2x positive if available)
 - Spans with entity, containing string of required type
- Train logistic regression

Extract fillers

- Search for mentions of target entity in collection
30w **entity** 30w
- Run NER to select potential fillers
- Run each of the classifiers
- For each accepted entity – filler pair, count and average classifier weights
- For each entity slot:
 - If single-valued, return top-scoring filler
 - If multiple-valued, return 5 top-scoring fillers
- Link fillers to entities using LNRM dictionary method

Results and conclusions

	SF-average
Best	77.9
Median	46.1
Stanford_UBC3	37.3
Stanford_UBC1	35.5

- 1 – Basic system
- 2 – Bug, same as 1
- 3 – Same as 2 but with more negative samples.

- Below median
- Premature version of the baseline system
 - Too liberal (few NILs)
 - Non-NILs over median
 - Filler in more than one slot

Thank you!

