Relation Extraction from Wikipedia Text

Wikipedia Mining Seminar
Universität des Saarlandes

Miriam Käshammer

January 18, 2010

Outline

Introduction

PORE: Positive-Only Relation Extraction from Wikipedia Text

Unsupervised Relation Extraction from Wikipedia

Relation Extraction

Definition

Automated or human-assisted acquisition of relations between concepts from textual or other data.

http://www.lt-world.org

- Subtask of Information Extraction
- Used for Database/Ontology population, Semantic Web annotations
- Traditional supervised machine learning approaches: annotated training data $\longrightarrow$ substantial human effort
- Need for RE algorithms that operate as unsupervised as possible

Relation Extraction

PORE: Positive-Only Relation Extraction from Wikipedia Text
G. Wang, Y. Yu and H. Zhu, ISWC, 2007

Unsupervised Relation Extraction by Mining Wikipedia Texts Using Information from the Web
Y. Yan, N. Okazaki, Y. Matsuo, Z. Yang and M. Ishizuka, ACL, 2009

Outline

Introduction

PORE: Positive-Only Relation Extraction from Wikipedia Text

Unsupervised Relation Extraction from Wikipedia

PORE

- Use the structure of Wikipedia articles to semi-automatically extract semantic relations from free Wikipedia text
- Core algorithm: B-POL (Bootstrapping positive-only learning)

PORE

- Use the structure of Wikipedia articles to semi-automatically extract semantic relations from free Wikipedia text
- Core algorithm: B-POL (Bootstrapping positive-only learning)

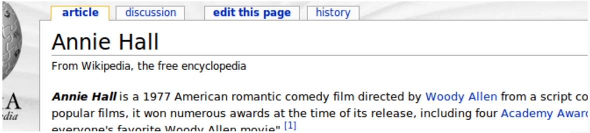
Steps

1. Extract **entity features** from semi-structured data of Wikipedia
2. Extract **context features** from the co-occurrence of two entities in one sentence in the Wikipedia text
3. For each relation, **filter** out irrelevant pairs
4. Conduct relation **classification** on the filtered set of pairs using B-POL

Entity Feature Extraction

Entity features describe Wikipedia entities (entries).

- **Definition features:** the head word of the first base noun phrase following a *be*-verb
  $\longrightarrow$ `film`, `comedy_film`

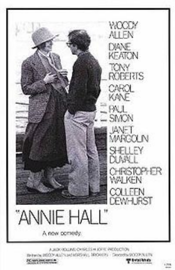


article | discussion | edit this page | history

Annie Hall

From Wikipedia, the free encyclopedia

Annie Hall is a 1977 American romantic comedy film directed by Woody Allen from a script co
popular films, it won numerous awards at the time of its release, including four Academy Awar
everyone's favorite Woody Allen movie" [1]



Annie Hall

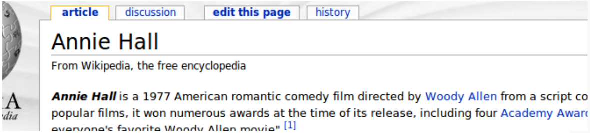
Directed by	Woody Allen
Produced by	Charles H. Joffe

Categories: English-language films | 1977 films | 1970s romantic comedy films | American films |
Actress Academy Award winning performance | Films set in Brooklyn | Films set in New York City
director won the Best Director Academy Award | Films whose writer won the Best Original Screer

Entity Feature Extraction

Entity features describe Wikipedia entities (entries).

- **Definition features:** the head word of the first base noun phrase following a *be*-verb
  $\longrightarrow$ `film`, `comedy_film`
- **Category features:** the head word of the first base noun phrase in each category phrase
  $\longrightarrow$ `film`, `comedy_film`, ...

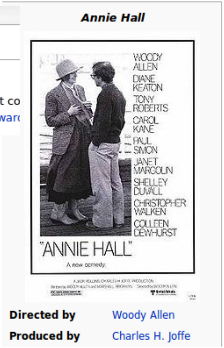


article | discussion | edit this page | history

Annie Hall

From Wikipedia, the free encyclopedia

Annie Hall is a 1977 American romantic comedy film directed by Woody Allen from a script co
popular films, it won numerous awards at the time of its release, including four Academy Awar
everyone's favorite Woody Allen movie" [1]



Categories: English-language films | 1977 films | 1970s romantic comedy films | American films |
Actress Academy Award winning performance | Films set in Brooklyn | Films set in New York City
director won the Best Director Academy Award | Films whose writer won the Best Original Screer

Entity Feature Extraction

Entity features describe Wikipedia entities (entries).

- **Definition features:** the head word of the first base noun phrase following a *be*-verb
  → `film`, `comedy_film`
- **Category features:** the head word of the first base noun phrase in each category phrase
  → `film`, `comedy_film`, ...
- **Infobox features:** predicate names from the infoboxes (white spaces replaced by underscores)
  → `directed_by`, `produced_by`, ...

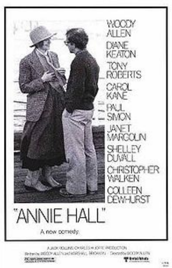
article | discussion | edit this page | history

Annie Hall

From Wikipedia, the free encyclopedia

Annie Hall is a 1977 American romantic comedy film directed by Woody Allen from a script co
popular films, it won numerous awards at the time of its release, including four Academy Awar
everyone's favorite Woody Allen movie" [1]

Annie Hall



Directed by	Woody Allen
Produced by	Charles H. Joffe

Categories: English-language films | 1977 films | 1970s romantic comedy films | American films |
Actress Academy Award winning performance | Films set in Brooklyn | Films set in New York City
director won the Best Director Academy Award | Films whose writer won the Best Original Screer

Context Feature Extraction

Context features describe co-occurrence contexts of pairs of Wikipedia entities in a sentence.

In the film "Heavenly Creatures", directed by Peter Jackson, Juliet Hulme had TB, and her fear of being sent ...

on the page *Tuberculosis in popular culture*

- Six entity pairs (`SUBJ`, `OBJ`): e.g. (`Heavenly Creatures`, `Peter Jackson`)
- For each pair, tokens in the left context, in the right context and tokens between the two entities are encoded as the context features

Filtering

- Very large number of entity pairs
- Use the entity features for filtering the pairs
- Features scoring function:

$$score(f) = |P_f| \times log(|C|/|C_f|)$$

C: complete set of pairs	C_f: set from C containing f
P: positive set in C ("seeds")	P_f: set from P containing f

- Score the entity features at each argument position (subject or object) and select the top k features ($k = 15$).
- Keep pairs in which entity features of SUBJ intersect with the *Salient Subject Features* and entity features of OBJ intersect with the *Salient Object Features*. $\Rightarrow C'$
- $U = C' - P$

Positive-only binary classification

Given:

- a collection C of context feature vectors of entity pairs
- a relation type R
- a set of positive training data $P \subset C$ (“seeds”)

Task: classify the unlabeled set $U = C - P$ into entity pairs which are of type R (positive set) and entity pairs which are not of type R (negative set)

B-POL

B-POL builds on top of the POL (positive-only learning) approach.

POL initially identifies very strong negative examples from the unlabeled data and then iteratively classifies more negative data until no such data can be found.

POL(P,U)

1. Use a weak classifier (Rocchio) to classify using P and U.
   $P_0 \leftarrow$ the data in U classified as positive; $N_0 \leftarrow U - P_0$
2. $N \leftarrow \varnothing; i \leftarrow 0$
3. Do loop until $N_i = \varnothing$
   $N \leftarrow N \cup N_i$
   Use v-SVM to classify P_i with P and $N \Rightarrow N_{i+1}$, P_{i+1}
   $i \leftarrow i + 1$
4. $P_u \leftarrow P_i$; return P_u

B-POL cont.

When P is very small $\Rightarrow$ low recall using POL

Extension of POL:
Add the newly generated positive data P_u identified by POL to the set of positive training samples and invoke POL again to generate more positive data. (bootstrapping)

B-POL(P,U)

1. $P_u \leftarrow \varnothing; i \leftarrow 0$
2. Do loop
   $i \leftarrow i + 1$
   ${P_u}^{(i)} \leftarrow$ positive examples returned from POL($P \cup P_u$,U)
   $P_u \leftarrow P_u \cup {P_u}^{(i)}; U \leftarrow U - {P_u}^{(i)}$
   Repeat until ${P_u}^{(i)} = \varnothing$
3. Return P_u

Evaluation

- Definitions of relations and the corresponding training instances are taken from the infoboxes of Wikipedia.
- 10.000 randomly selected Wikipedia pages (no disambiguation or list-of pages) $\Rightarrow$ ~130.000 pairs of entities
- Construction of a gold standard set for each relation tested

Table 1. Information about the four relations.

Relation	Source	#GS	#U	#(GS ∩ U)
album-artist	album_infobox#artist	274	392	260
film-director	infobox_movie#director	121	286	115
university-city	infobox_university#city	74	208	71
band-member	infobox_band#current_members	117	477	103

Recall $\frac{\#(GS \cap U)}{\#GS}$ at this stage is relatively high.

Evaluation cont.

Extraction performance of B-POL

#P	method	album-artist	film-director	university-city	band-member
		P/R/F1	P/R/F1	P/R/F1	P/R/F1
40	*T-POL'*	96.7/36.5/47.8	82.8/50.6/60.6	65.4/74.4/**68.6**	70.2/25.0/35.7
	T-POL	89.6/49.8/59.2	82.2/58.2/66.4	62.0/76.8/68.1	67.6/25.0/34.8
	B-POL	86.6/77.5/**79.9**	69.4/81.2/**73.2**	47.2/84.8/58.5	46.8/57.6/**47.1**
	M-SVM	93.6/40.4/54.5	71.2/32.8/41.4	17.4/36.9/19.5	35.4/29.7/ 27.5
30	*T-POL'*	97.4/45.8/58.8	85.5/51.1/62.2	75.1/67.7/70.5	74.3/24.5/35.9
	T-POL	93.2/56.7/68.2	83.7/51.0/61.8	70.7/72.6/**70.6**	67.6/22.0/32.4
	B-POL	90.6/70.2/**76.5**	73.4/69.6/**68.6**	62.7/79.0/68.5	58.5/46.6/**49.3**
	M-SVM	93.4/46.2/58.0	72.1/37.9/44.8	20.9/33.7/21.9	36.1/32.5/30.0
20	*T-POL'*	97.1/34.6/48.0	84.6/37.7/49.9	80.3/63.6/70.5	77.7/21.7/33.5
	T-POL	93.5/52.8/63.7	81.3/47.0/56.5	79.8/64.0/70.2	72.3/21.0/31.5
	B-POL	90.0/69.2/**76.4**	74.7/64.1/**66.6**	75.3/70.1/**71.6**	67.9/32.3/**41.9**
	M-SVM	93.8/42.4/55.9	73.1/40.5/46.9	27.0/31.6/26.0	39.4/32.9/29.8
10	*T-POL'*	99.1/35.3/50.7	89.1/32.1/45.7	82.5/57.7/66.7	81.4/12.5/21.2
	T-POL	96.7/40.5/53.8	86.2/30.5/42.5	84.1/54.1/64.8	76.7/15.2/24.6
	B-POL	95.0/48.6/**61.3**	83.2/41.3/**51.0**	82.7/58.1/**67.5**	74.0/19.9/**30.1**
	M-SVM	93.4/46.3/58.9	78.3/31.4/42.7	32.1/28.1/29.1	40.6/32.8/26.4

Outline

Introduction

PORE: Positive-Only Relation Extraction from Wikipedia Text

Unsupervised Relation Extraction from Wikipedia

An integrated Approach

Dependency patterns from dependency analysis

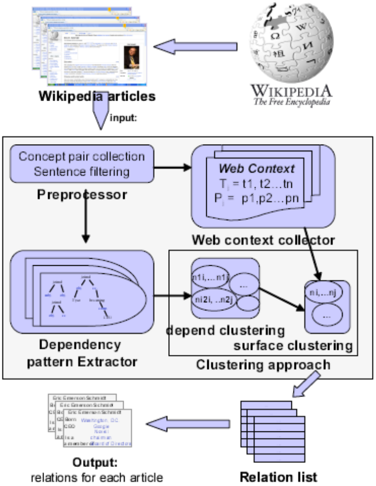
- Linguistic technologies to abstract away from different surface realizations of semantic relations
- Expected to be more accurate ⇒ good precision
- Dependency parsing requires text of good quality
- **Wikipedia** as a local corpus

Surface patterns generated from redundant Web information

- Contribute greatly to the coverage
- The **Web** as a global corpus

⇒ Bridge the gap separating "deep" linguistic technology and redundant Web information for IE tasks

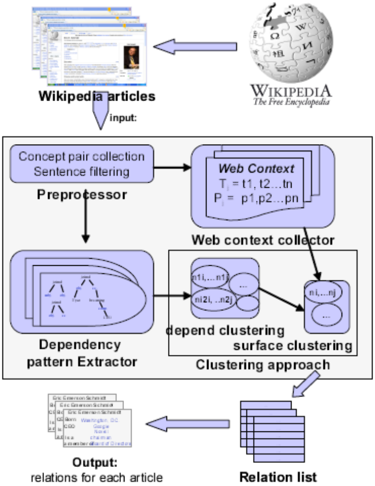
The framework



Assumption: It is likely that a salient semantic relation r exists between a page p and a related page p' which is linked on page p.

$\Rightarrow$ Relation Extraction between the entitled concept (ec) and a related concept (rc), which appear as links in the text of the article.

The framework



Assumption: It is likely that a salient semantic relation r exists between a page p and a related page p' which is linked on page p.

$\Rightarrow$ Relation Extraction between the entitled concept (ec) and a related concept (rc), which appear as links in the text of the article.

Given: a set of Wikipedia articles

Output: a list of concept pairs for each article with a relation label assigned to each concept pair

Preprocessor

- Split the text of an article into sentences.
- Select sentences containing one reference of the entitled concept and one of a linked concept
  $\Rightarrow$ a set of concept pairs, each associated with a sentence

Dependency Pattern Extractor

1. Parse the sentence for a concept pair; induce the shortest dependency path with the entitled concept and the related concept.
2. Generate sub-paths of the shortest path as dependency patterns (frequent tree-mining algorithm, Zaki, 2002).

Web Context Collector

- Query a concept pair using a search engine (Google).
- Extract two kinds of relational information from the retrieved snippets:
    1. Ranked relational terms (as keywords)
    2. Surface patterns

Web Context Collector cont.

Relational term ranking

- Relational terms are defined as verbs and nouns in the snippet.
- For each concept pair, collect a list of relational terms.
- Rank the relational terms of all concept pairs (entropy-based algorithm, Chen et al., 2005) $\Rightarrow T_{all}$
- For each concept pair, sort the list T_{cp} according to the terms' order in T_{all}
- For each concept pair, select the top term as a keyword t_{cp}

Web Context Collector cont.

Surface pattern generation

- Consider a snippet sentence
- Content Words (`CW`): entitled concept, related concept and the keyword t_{cp}
- Functional Words (`FW`): verbs, nouns, prepositions and coordinating conjunctions
- General form of surface patterns:
  `CW1 Infix CW2 Infix CW3` with `Infix = FW*`
  e.g. `ec` *assign* `rc` *as* `ceo`
  `ceo` *of* `ec rc`

Clustering

k-Means Algorithm

Initial Centroid Selection based on the keyword t_{cp} of each concept pair:

- Group all concept pairs by their keyword t_{cp}
- $G = \{G_1, G_2, ...G_n\}$ with each $G_i = \{cp_{i1}, cp_{i2}, ...\}$ being a group of concept pairs that share the same keyword
- Rank the groups by their number of concept pairs and choose the top k groups (stability-based criteria, Chen et al., 2005)
- For each group G_i, select a centroid c_i:

$$c_i = \arg\max_{cp \in G_i} |\{cp_{ij} | (dis_1(cp_{ij}, cp) + \lambda * dis_2(cp_{ij}, cp)) \leq D_z, 1 \leq j \leq |G_i|\}|$$

Clustering cont.

Dependency pattern distance

$$dis_1(cp_i, cp_j) = 1 - \frac{|DP_i \cap DP_j|}{\sqrt{|DP_i| * |DP_j|}}$$

DP_x: dependency pattern set of the concept pair cp_x

Clustering cont.

Surface pattern distance

$dis_2(cp_i, cp_j)$

Input: $SP_1 = \{sp_{11}, ..., sp_{1m}\}$, $SP_2 = \{sp_{21}, ..., sp_{2n}\}$

1. Define a $m \times n$ distance matrix A:
$$\{A_{ij} = \frac{LD(sp_{1i}, sp_{2j})}{Max(|sp_{1i}|, |sp_{2j}|)}; 1 \leq i \leq m; 1 \leq j \leq n\}$$
2. $dis \leftarrow 0$
3. for $min(m, n)$ $times$ do
$(x, y) \leftarrow \arg\min_{0<i<m;0<j<n} A_{ij}$
$dis \leftarrow dis + A_{xy}/min(m, n)$
$A_{x*} \leftarrow 1; A_{*y} \leftarrow 1$
4. return dis

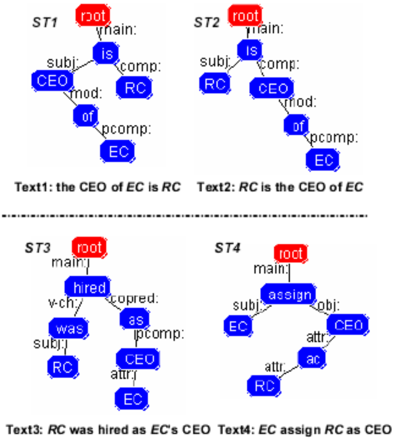
Depend Clustering

- Given the initial k centroids, merge the concept pairs into k clusters according to their dependency patterns.
- Each concept pair cp_i has a set of dependency patterns DP_i.
- Distance between two concept pairs cp_i and cp_j: $dis_1(cp_i, cp_j)$

Steps:

1. Assign each concept pair to the cluster with the closest centroid if the distance is smaller than D_l.
2. Then recompute each centroid based on the current members of the cluster.
3. Repeat steps 1. and 2. until the centroids do not change anymore.

CEO-relation



Surface Clustering

- Merge more concept pairs into the existing clusters using the surface patterns to improve the coverage.
- Each concept pair cp_i has a set of surface patterns SP_i.
- Distance between two concept pairs cp_i and cp_j: $dis_2(cp_i, cp_j)$

Steps:

1. Assign each concept pair (which has not yet been assigned to a cluster by depend clustering) to the cluster with the closest centroid if the distance is smaller than D_g.
2. Then recompute each centroid based on the current members of the cluster.
3. Repeat steps 1. and 2. until the centroids do not change anymore.

Clustering cont.

Result

1. k clusters of concept pairs
2. Use the centroid pair to assign a label to the corresponding relation

Evaluation

Comparison with another method

- 526 articles from the Wikipedia category "American chief executives"
- 7310 concept pairs
- $k = 18$
- 15 clearly identifiable relations
- `#Ins.`: number of concept pairs in the cluster
- `pre`: precision of the cluster
- coverage $= \dfrac{\#\text{correct Ins.}}{\#\text{all concept pairs}}$

method	Existing method (Rosenfeld et al.)		Proposed method (Our method)	
Relation (sample)	# Ins.	pre	# Ins.	pre
chairman (*x be chairman of y*)	434	63.52	547	68.37
ceo (*x be ceo of y*)	396	73.74	423	77.54
bear (*x be bear in y*)	138	83.33	276	86.96
attend (*x attend y*)	225	67.11	313	70.28
member (*x be member of y*)	14	85.71	175	91.43
receive (*x receive y*)	97	67.97	117	73.53
graduate (*x graduate from y*)	18	83.33	92	88.04
degree (*x obtain y degree*)	5	80.00	78	82.05
marry (*x marry y*)	55	41.67	74	61.25
earn (*x earn y*)	23	86.96	51	88.24
award (*x won y award*)	23	43.47	46	84.78

Evaluation

Contribution of the different patterns

Pattern type	#Instance	Precision	Coverage
dependency	1127	84.29	13.00%
surface	1510	68.27	14.10%
Combined	2314	75.63	23.94%

Conclusion

Two Approaches to Relation Extraction

1. **Semi-supervised**: seed-based bootstrapping approach
   - Uses the structure of Wikipedia to extract entity features and entity pairs with context features
   - Positive-only learning
   - Finds instances of one specified relation in Wikipedia
2. **Unsupervised**: clustering approach
   - Uses the link structure of Wikipedia to extract entity pairs in a sentence
   - Combines linguistic analysis with redundancy information from the Web
   - k-means clustering
   - Finds the major relations and the corresponding instances in the given corpus (e.g. Wikipedia articles)

References

Wang, G., Yu, Y., and Zhu, H. (2007).
PORE: Positive-Only Relation Extraction from Wikipedia Text.
The 6th International Semantic Web Conference (ISWC).

Yan, Y., Matsuo, Y., and Ishizuka, M. (2009a).
An Integrated Approach for Relation Extraction from Wikipedia Texts.
CAW2.0.

Yan, Y., Okazaki, N., Matsuo, Y., Yang, Z., and Ishizuka, M. (2009b).
Unsupervised Relation Extraction by Mining Wikipedia Texts Using Information from the Web.
Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 1021–1029.