

Related Entity Finding: University of Waterloo at TREC 2010

Olga Vechtomova (ovechtom@uwaterloo.ca)

Overview

The University of Waterloo participated in the Related Entity Finding (REF) task of the Entity track. Our goal is to investigate whether related entity finding problem can be addressed by unsupervised approaches that rely primarily on statistical methods and common linguistic tools, such as named-entity taggers and syntactic parsers. We approach the related entity finding problem by first retrieving documents in response to the query, and extracting an initial set of candidate entities from the text of the documents. As a separate step, we automatically construct a set of seed entities, which represent hyponyms of the target entity category specified in the narrative, and then rank the candidate entities by their similarity to the seeds.

Stage 1: Extract entities

- Construct queries from the entity_name and narrative sections of topics
- Retrieve top 50 web pages using a commercial search engine
- Find snippets containing query terms
- Tag them using a Named Entity (NE) tagger (Ratinov and Roth, 2009)
- Retain entities with NE tags corresponding to the target entity type given in the topic
- Rank entities by TF*IDF (run **UWAT1**)

Topic example:
 <query><num>23</num>
 <entity_name>The Kingston Trio</entity_name>
 <entity_URL>clueweb09-en0009-81-29533</entity_URL>
 <target_entity>organization</target_entity>
 <narrative>What recording companies now sell the Kingston Trio's songs? </narrative></query>

Stage 4: Compute distributional similarity between seeds and candidates

- For each entity (seed and candidate):
- Retrieve 200 documents using BM25
- Retain sentences containing the entity
- Do syntactic parsing with Minipar (Lin, 1993)
- Extract grammatical dependency triples containing the entity
- Example dependency triples containing the seed entity "Capitol Records":
 - Capitol Records N:nn:N label
 - release V:subj:N Capitol Records
 - album N:nn:N Capitol Records

- Transform a dependency triple into a feature by removing the entity:
 - X N:nn:N label
 - release V:subj:N X
 - album N:nn:N X

- A vector for each entity consists of features and their frequencies
- Only triples co-occurring with at least 50% of seed entities are used as vector features
- Compute similarity between the vectors of each candidate and seed using BM25 with query weights (Spärck Jones et al., 2000):

$$QACW_{c,s} = \sum_{f=1}^F \frac{TF(k_f+1)}{K+TF} QTF \cdot IDF_f \quad (1)$$

- The IDF of a feature (IDF_f) is approximated as IDF of the word it contains:

release V:subj:N X

Entity track topic

1. Extract and rank candidate entities from top 50 retrieved documents

2. Extract category name from narrative

category name

3. Extract hyponyms of the category name from the Web and select seed entities from them

seed entities

4. Compute distributional similarity between candidate and seed entities

5. Rank candidate entities by similarity to all seed entities

Stage 2: Extract target entity category name from narrative

- Do POS tagging and NP-chunking of the topic narrative
- Apply a set of rules to extract the category name of the target entities.

Example:
 <narrative>What recording companies now sell the Kingston Trio's songs? </narrative>
 [WhatWP] [recordingNN companiesNNS] now/RB sell/VBP [theDT Kingston/NNP Trio's/NNP songs/NNS] ?/.

Category name: recording companies

Stage 3: Find seeds

- Find hyponyms of the category name
- Construct queries using Hearst's (1992) hyponym patterns:
 - "recording companies such as"; "such recording companies as"; "or other recording companies"; "and other recording companies"; "recording companies including"; "recording companies especially"
- Submit them to a search engine
- Do NE-parsing of the retrieved sentences containing the pattern:
 - "In large recording companies such as [ORG EM] , the mastering process was usually controlled by specialist staff technicians who were conservative in their work practices ."
- Extract NEs with the correct NE tag
- Use as seeds only those NEs that exist in the list of candidate entities output by Stage 1

- Seeds extracted for the category name "recording company":
- warner bros
 - decca
 - columbia records
 - capitol records
 - bear family records

Stage 5: Rank candidate entities

- Calculate candidate's similarity to all seeds weighted by the seed's TF*IDF:

$$EntitySeed_c = \sum_{s=1} TFIDF_s \times QACW_{c,s} \quad (2)$$

- Calculate the final score for each candidate (run **UWAT2**):

$$TFIDFEntitySeed_c = \beta \times TFIDF_c + (1 - \beta) \times EntitySeed_c \quad (3)$$

Results

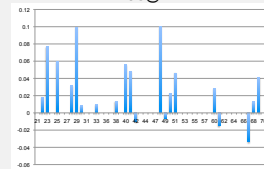
Run	nDCG@R	P10	MAP	RPrec	Rel Retr.	Pri. Retr.
UWAT1 (TF*IDF)	0.1264	0.0957	0.0608	0.1033	95	151
UWAT2 (TFIDFEntitySeed)	0.1393* (10.2%)	0.1106 (15.6%)	0.0722* (18.8%)	0.1223 (18.4%)	96	154

* and ** are statistically significant at 0.05 and 0.01 levels (2-tailed paired t-test)

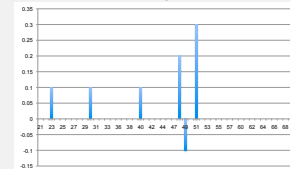
21 topics in UWAT1 and 25 in UWAT2 are above the task median in NDCG@R (number of judged topics: 47)

Differences between UWAT1 and UWAT2

nDCG@R



P10



Examples of top 10 retrieved entities

Topic 23: What recording companies now sell the Kingston Trio's songs?

UWAT1	UWAT2
kingston trio	kingston trio
kingston trio on record	capitol
new kingston trio	capitol records
purple onion	decca records
capitol	beach boys
capitol records	columbia records
the kingston	new christy minstrels
kingston trio story	fleetwood mac
queue	mta
the guard years	elektra records

UNIVERSITY OF
WATERLOO