

A Web-oriented Language Modeling Approach for Question Related Entity Finding

Ludovic Bonnefoy, Patrice Bellot and Michel Benoit

iSmart - Université d'Avignon (LIA)
Michel Benoit - Patrice Bellot

19 novembre 2010

TREC 2010 Related Entity Finding topic

```
<query>  
  <entity\_name>Smithsonian Institution</entity\_name>  
  <entity\_URL>clueweb09-en0011-99-06195</entity\_URL>  
  <target\_entity>person</target\_entity>  
  <narrative>Find the members of the Board of Regents of the  
  Smithsonian Institution.</narrative>  
</query>
```

TREC 2007 QA topic

```
<target id = "220" text = "International Management Group (IMG)">  
  <q id = "220.4" type="LIST">  
    Who are members of the board of the IMG?  
  </q>  
</target>
```

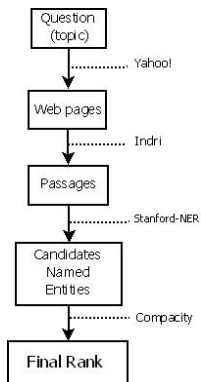
REF topic in QA form

```
<target id = "220" text = "Smithsonian Institution">  
  <q id = "220.4" type="LIST">  
    Find the members of the Board of Regents of the Smithsonian  
    Institution.  
  </q>  
</target>
```

QA topic in REF form

```
<query>  
  <entity\_name>International Management Group (IMG)</entity\_name>  
  <entity\_URL>clueweb09-en0001-64-06494</entity\_URL>  
  <target\_entity>person</target\_entity>  
  <narrative>Who are members of the board of the IMG?</narrative>  
</query>
```

Question-Answering approach



Passages

- ▶ One sentence length

Named Entity Recognition

- ▶ Pers, Org and Loc : Stanford-NER
- ▶ Products : Home-made rules

Compacity

- ▶ $Comp(EC_i) = \frac{1}{|QW|} \sum_{w \in QW} \frac{Z_j}{|R_j|+1}$
- ▶ Laurent Gillard and al : Relevance Measures for Question Answering, The LIA at QA@CLEF-2006



- ▶ Ex : ■ Keyword ■ NE ■ Other words

Observation

Observation

- ▶ QA systems deal with few types between 3-4 and 50
- ▶ because corpus are not available for machine learning approaches and rules are difficult to maintains.

Our goal

- ▶ To deal with any type of named entities (types as broad as *person* or as specific as *teammates* or *scotch whisky distilleries*)

Idea

- Specific words distribution in web pages related to a type

Example : *'portable mp3 player'*

Words with unusual frequency : mp3, player, headphones, capacity, format, ...

Idea

- ▶ Specific words distribution in web pages related to a type
Example : *'portable mp3 player'*
Words with unusual frequency : mp3, player, headphones, capacity, format, ...
- ▶ idem for a given candidate NE
Example : *'Winnie the Pooh'*
Specific words : fictional, character, bear, friends, disney, ...

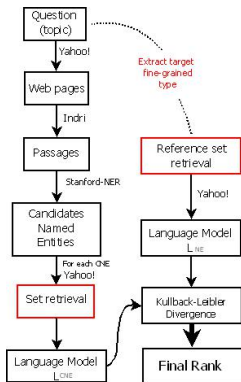
Idea

- ▶ Specific words distribution in web pages related to a type
Example : *'portable mp3 player'*
Words with unusual frequency : mp3, player, headphones, capacity, format, ...
- ▶ idem for a given candidate NE
Example : *'Winnie the Pooh'*
Specific words : fictional, character, bear, friends, disney, ...
- ▶ language model of a candidate NE close to the one of its types
Example : *'iPod'*
Words : apple, mp3, player, format, headphones, media, ...

Idea

- ▶ Specific words distribution in web pages related to a type
Example : *'portable mp3 player'*
Words with unusual frequency : mp3, player, headphones, capacity, format, ...
- ▶ idem for a given candidate NE
Example : *'Winnie the Pooh'*
Specific words : fictional, character, bear, friends, disney, ...
- ▶ language model of a candidate NE close to the one of its types
Example : *'iPod'*
Words : apple, mp3, player, format, headphones, media, ...
- ▶ ⇒ Measure the distance between language models of ECs from the target type's one

Candidate named entities categorisation



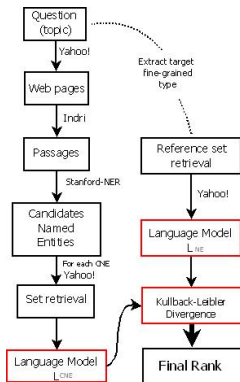
Fined-grained type extraction

- ▶ **Carriers** that Blackberry makes phones for.
- ▶ **Winners** of the ACM Athena award.
- ▶ **Airlines** that currently use Boeing 747 planes.

Web pages retrieval

- ▶ Querying Yahoo!
- ▶ Request : NE
- ▶ Retrieve the 10 top ranked ones

Candidate named entities categorisation(2)



Language models

- ▶ Unigrams
- ▶ Smoothing : Dirichlet

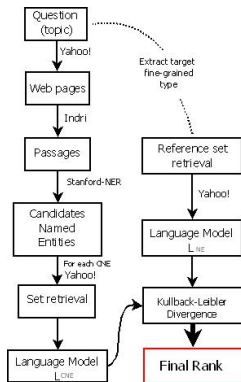
$$p(w|d) = \frac{\mu}{\sum_{w \in V} tf(w,d) + \mu} p(w|C)$$

Kullback-Leibler divergence

- ▶ Between target type's language model L_{NE} and candidate named entity's model L_{CNE}

- ▶ $DKL(L_{CNE}|L_{NE}) = \sum_i L_{CNE}(i) \log \frac{L_{CNE}(i)}{L_{NE}(i)}$

Final Ranking



Four ways

- ▶ Compacity only
- ▶ KL divergence only
- ▶ Harmonic mean of candidate NE rank by compacity and the one by KL divergence
- ▶ Multilayer perceptron :
 - ▶ Find an effective combination of compacity, KL divergence, passage's score and idf
 - ▶ Train on 45 TREC QA 2006-2007 topics

Homepages

- ▶ Query Yahoo! with : "NE + homepage"
- ▶ Filter pages according to their type (machine learning approach)
- ▶ Map to the ClueWeb09 id

Official results

	Comp	Div	RDC	LearnDPI	Best	Median
P@10	.0468	.0213	.0362	.0532 (+14%)	-	-
nDCG@R	.0737	.0428	.0610	.0766 (+4%)	≈.38	≈.12
map	.0261	.0129	.0200	.0305 (+17%)	-	-
Rprec	.0463	.0189	.0373	.0591 (+27%)	-	-

TABLE: Official evaluation of our QA system over TREC 2010 Entity track topics for Precision at 10 documents (P@10) and nDCG@R measures (Compacity only, K-L divergence only, Harmonic mean and Machine Learning approach compared to Best and Median official results)

Analyze

Training Topic, Entity Track 2009

```
<query>
<entity_name>Michael Schumacher</entity_name>
<entity_URL>http://www.michael-schumacher.de/?lang=uk</entity_URL>
<target_entity>person</target_entity>
<narrative>Michael's teammates while he was racing in Formula 1.</narrative>
</query>
```

nico rosberg
eddie irvine
felipe massa
rubens barrichello
johnny herbert
ross brawn
van diemen
j. lehto
david coulthard
muhammad ali

nico rosberg
felipe massa
muhammad ali
sebastian vettel
joe louis
toro rosso
fernando alonso
giancarlo fisichella

FIGURE: 10 top ranked NEs before

FIGURE: 10 top ranked NEs after

Future works

Entity 2011

- ▶ Looking for an effective way to find homepages in ClueWeb09
- ▶ Using knowledges bases (freebase, dbpedia, ...)
- ▶ Using Wikipedia
- ▶ Participate in ELC.

General Works

Identify and extract ontological characteristics of NEs

Thank you for your attention

Questions ?

ludovic.bonnefoy@ismart.fr

Web pages categorisation

Ressources

- ▶ SMO : SVM classifier
- ▶ 7genres : corpus with 200 web pages for 7 types
blog, shopping, faq, frontpage, personal homepage, listing, search page

Features

- ▶ Words and POS frequencies
- ▶ Tags frequencies
- ▶ Average length of sentences, documents (in words), ...

Performances

- ▶ 10 cross validation : 99,7%
- ▶ but..