LIA-iSmart at the TREC 2011 Entity track : Entity List Completion Using Contextual Unsupervised Scores for Candidate Entities Ranking

Ludovic Bonnefoy
LIA - iSmart
University of Avignon
Avignon - Aix en provence, France
ludovic.bonnefoy@etd.univ-avignon.fr

Patrice Bellot
LSIS
Aix-Marseille University
Marseille, France
patrice.bellot@lsis.org

Abstract

This paper describes our participation in the Entity List Completion (ELC) task at Entity track 2011. Our approach combined the work done for the Related Entity Finding 2010 task with some new criteria as the proximity or the similarity between a candidate answer with the correct answers given as examples or their cooccurrences.

1 Introduction

The aim of the Entity track is to evaluate entity-related searches on Web data. The third edition of the track features two main tasks (REF and ELC) and a pilot task (REF-LOD) (Balog et al., 2011).

The *Related Entity Finding* task (REF) is formulated as follows :
"Given an input entity, by its name and homepage, the type of the target entity, as well as the nature of their relation, described in free text, find related entities that are of target type, standing in the required relation to the input entity".
As defining entities on the Web is still an unsolved problem, it was decided to represent entities by their homepage URL, used as unique identifier. These URLs have to be extract from the English portion of ClueWeb09[1] which contains approximately 500 million pages. We participated to this task last year with results just under the median (Bonnefoy et al., 2010).

The REF-LOD task has the same definition as the REF task but the unique identifiers are URIs from the Linked Open Data (LOD) sample provided by the Sindice's team.

The *Entity List Completion* task (ELC) was a pilot task last year and is now one of the two main tasks. There is two differences with REF-LOD:

- A number of relevant entities (homepages and the name if available) are given in the topic definition, as examples,
- In addition to a broad type of the target entities there is a more specific type from the DBPedia Ontology.

This year we decided to participate to the *Entity List Completion* task in order to explore the impact of the use of previous results that we are confident in (ie. the examples) to (re)rank other candidates.

As said above, in 2010 we participated to the REF task and we decided that for 2011 ELC task we were going to reuse the core of what we implemented and add new criteria, in order to use the information given by the examples, to rank candidate answers associated with their URI.

The paper is broken down as follows : we briefly describe how we extract candidate answers and in second time we describe all the criteria used and how we combined them for each run.

2 Finding candidate answers

The first thing to do is to be able to find candidate answers (named entities) from the topic in input. To do this we reused what we did last year and presented in Fig.1.

First, we had to build a query Q in order to retrieve relevant documents. We build it by concatenating the source entity to the words of the narrative field (more sophisticated approaches, like only using the commons and proper nouns of the narrative, seem to be less effective).

For the Web runs, the query was used to retrieve a set of related web pages by querying the

[1] http://boston.lti.cs.cmu.edu/Data/clueweb09/

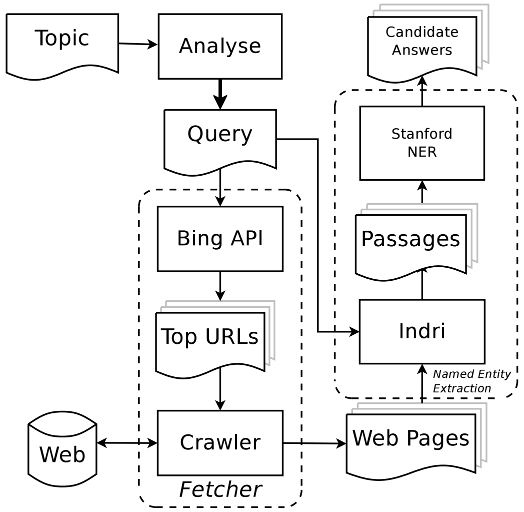

Fig. 1. Candidate answers finder (CAF)

web search engine Bing. In 2010 we used Boss (Yahoo!'s API) but this service is now non free. The 100 top ranked web pages were downloaded, cleaned of HTML tags and parsed in sentences.

For the obligatory run using the ClueWeb09 collection we indexed the ClueWeb collection with Indri[2] and used the embedded stoplist along with the standard Krovetz stemmer. We queried it and the 100 top ranked web pages were then extracted from the Warc files, cleaned of HTML and parsed in sentences too.

For all the runs, the sentences were then indexed with Indri. Finally, we queried Indri with Q and kept the 500 top ranked passages.

Lastly, candidate named entities were extracted from this set of passages by using the Stanford-NER[3] and some homemade rules.

3 Ranking candidate answers

The next step deals with candidate named entities ranking.

3.1 Compacity

The first criterion we used to rank the candidates named entities is the "Compacity" score (Gillard et al., 2006) and that we already used in 2010. It measures the density of the query words in a passage around a given candidate entity (a correct answer to a query tends to appear in the texts near of the query words). Compacity is defined as :

$$Compacity(E, P) = \frac{1}{|QW|} \sum_{w \in QW} \frac{Z_w}{R_w + 1} \qquad (1)$$

with QW being the set of query words (elements extracted from the topic to get the web pages), $|QW|$ the cardinality of this set and w one of them. Let E be a candidate named entity, R_w the distance (in number of words) between w and the candidate named entity in the passage P. Let Z_w be the number of query words between w and the E (both included).

3.2 An unsupervised measure of what extent a named entity is of a given type or is close to an other entity

Last year we tried to find a way to determine to what extent candidate answers to a natural language question may be associated to a given type of entity and how we can use this information to rank them. Our goal was to be able to deal with any type of entities as broad as "person" or as specific as "scotch whiskey distilleries".

Our idea, inspired by the distributional hypothesis (Sahlgren, 2008), that seems to work relatively well (Bonnefoy et al., 2011), is that we could do it by comparing the words distribution in web pages related to an entity to the one in web pages related to a given type :

- Obtain a first set of web pages related to the type, by querying a web search engine with the type (e.g.: "*science-fiction writers*"). This set is called "reference set". Obtain a second set, related to the entity, by querying the web search engine with it (e.g. : "*Isaac Asimov*").
- Compute, for each set, its words distribution (Dirichlet smoothing) :

  $$p'(w|s) = \begin{cases} p_s(w|s) \text{ if w is in the set} \\ \alpha_d p(w|C) \text{ otherwise} \end{cases} \qquad (2)$$

  where $p'(w|s)$ is the probability of word w in the set S, $p_s(w|s)$ is the smoothed probability of w, $p(w|C)$ the Laplace smoothed probability of w in a collection C (consists of 10% of the ClueWeb09 corpus) and α_d is a multiplier. $p_s(w|s)$ and α_d are estimated as:

  $$p_s(w|s) = \frac{tf(w,s) + \mu.p(w|C)}{\sum_{w' \in V} tf(w',s) + \mu} \qquad (3)$$

  $$\alpha_d = \frac{\mu}{\sum_{w \in V} tf(w,s) + \mu} \qquad (4)$$

  where $tf(w,s)$ is the term frequency of w in the set s, V is the set of all words w' in s and μ is a multiplier with a value set to 2000 (according to (Chen and Goodman, 1996) for newspapers and largest collection).

[2] http://www.lemurproject.org/indri/
[3] http://nlp.stanford.edu/ner/index.shtml

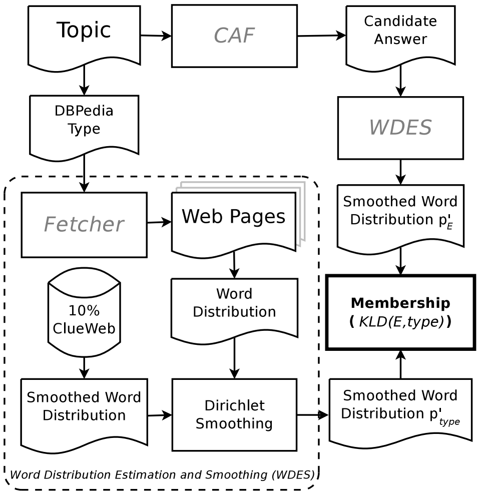
Topic
CAF
Candidate Answer
DBPedia Type
WDES
Fetcher
Web Pages
Smoothed Word Distribution p'_E
10% ClueWeb
Word Distribution
Membership (KLD(E,type))
Smoothed Word Distribution
Dirichlet Smoothing
Smoothed Word Distribution p'_type
Word Distribution Estimation and Smoothing (WDES)

Fig. 2. Measure of the membership of a named entity to a given type.

- Compare the words' probability p'_E, in documents associated to the entity, to the reference one p'_{type}, associated to the type. For this, we compute the Kullback-Leibler divergence between them :

$$KLD(E, type) = \sum_i p'_E(i).log\frac{p'_E(i)}{p'_{type}(i)} \quad (5)$$

where $KLD(E, type)$ is the Kullback-Leibler divergence for the given entity E and the type, $p'_E(i)$ (resp. $p'_{type}(i)$) is the probability of the i^{th} word in documents associated to the entity E (resp. to the type).

The degree of membership of a named entity to a type (see Fig2.) is then defined as :

$$Membership(E, type) = KLD(E, type) \quad (6)$$

And the degree of similarity of an entity to an other one (see Fig.3) is then defined as :

$$Similarity(E, E') = KLD(E, E') \quad (7)$$

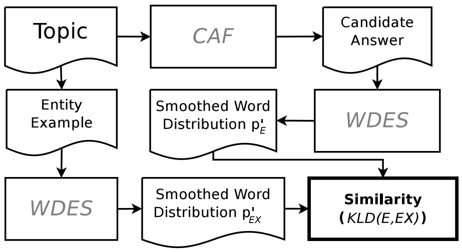

Fig. 3. Measure of what extent a named entity is similar to an other one.

3.3 Does the candidate answers cooccur with examples?

We thought that if a candidate named entity occurs in documents in which examples occur too, it must be favored in comparison to candidate answers that do not. We propose three ways for doing this.

3.3.1 Documents cooccurrences

The first one depends of how many times a candidate named entity is in a document where there is examples and how many of them are presents. This is formulated as :

$$SD(E) = \sum_{i \in D} \frac{\ln(x_i + n_i + 1)}{n_i} \quad (8)$$

where D is the set of the 100 web pages, x_i is either 0 (if E does not occur in D_i) or equal to the number of unique examples in D_i, n_i is the number of unique entities (including examples) in D_i. We choose to use a logarithm in order to give an important advantage to a candidate answer which cooccurs with some examples compared to one that doesn't but also in order to give only a slight advantage to a candidate answer which cooccurs with a huge number of examples over than one that occurs with a few number of them. The $n_i + 1$ in the logarithm allows to obtain results strictly positive.

3.3.2 Cooccurrences in homepage's website lists and tables

Our two others propositions to exploit cooccurrences of candidate answers and examples are using lists and tables in source entity homepage's website. As said above there is in the topic, in addition to the name of the source entity, the url of its homepage. By looking at the homepage of named entities from the 2009 topics, we noticed that some sub-pages contain all the correct answers, most of the time in lists or tables. Moreover, the url or the title of these web pages often contains either the type of named entity we are looking for or some words of the narrative.
For example we can consider the 16th topic which has for source entity *"Mancuso Quilt Festivals"*, for homepage *"http://www.quiltfest.com/"* and for target entity type *"sponsors"*. One of the web pages of this website is *"http://www.quiltfest.com/sponsor.asp"* and it contains all the correct answers to the topic (their name and a link to their homepage) in a table.

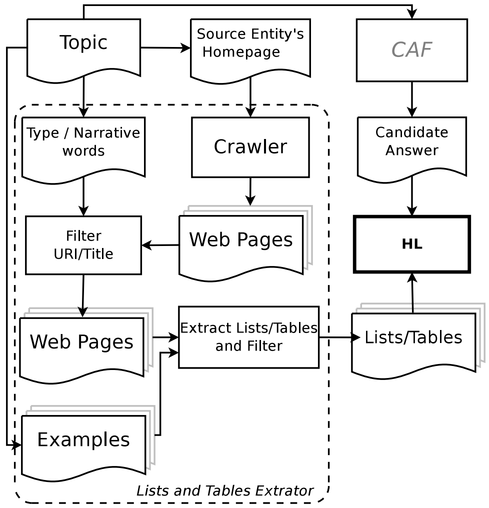

Fig. 4. Measuring cooccurrences in homepage's website's lists and tables.

Here's how we proceeded (see Fig. 4) :

First we crawled the homepage's website and kept all the web pages which have the same root (ie. which start with *"http://www.quiltfest.com/"* for the previous example) and which have some of the target type words in their url or title (we then refer to this way under the name "HLT") or some of the narrative (non-stop)words (this other way is called "HLN").

We then extracted all the lists and tables from the web pages kept and we discarded all the ones without one example at least. If for a topic, we did not have one list or table at least, this metric was not used for this one. In the other case, a score is associated to each candidate named entity (same formula for HLT and HLN) :

$$HL(E) = \sum_{i \in L} \frac{\ln(x_i + n_i + 1)}{n_i} \qquad (9)$$

where L is the set of lists and tables kept, x_i is either 0 (if E does not occur in L_i) or the number of unique examples in L_i +1 (to count the entity), n_i is the number of unique entities (including examples) in L_i.

3.4 Are the context of candidates and examples close?

We really think that if a candidate answer shares a same context with one or more examples then the candidate entity is probably a relevant answer too. We already propose in 3.2 a way to measure how much they share common vocabulary and with more or less the same distributions. Here we want to explore an other approach for doing this.

We retrieved for each example and each candidate for a topic the 100 top ranked snippets returned by Bing (with for query an example or a candidate answer). Then, we used a KMeans algorithm (the one includes in Mallet[4]) to assign in two clusters only all the snippets. The hope is that all the snippets corresponding to the examples are assigned to the same cluster and that all the snippets corresponding to relevant answers are assigned to this cluster too. To each candidate named entity we gave a "context score" (CS) according to :

$$CS(C) = \frac{(c_1 * \sum_{i \in E} e_{i,1}) + (c_2 * \sum_{i \in E} e_{i,2})}{c_1 + c_2} \qquad (10)$$

where C is a candidate answer, c_1 (resp. c_2) is the number of snippets corresponding to the candidate answer in the first cluster (resp. the second cluster), E the set of examples and $e_{i,1}$ (resp. $e_{i,2}$) is the number of snippets corresponding to the i_{th} example in the first cluster (resp. the second cluster). The denominator is for normalization if there are less than 100 snippets for a candidate answer.

3.5 Confidence in candidate answer's URI

To select correct URIs in the Sindice's LOD dump we looked for URIs which are subjects of RDF triples and satisfy the following constraints :

1. The name of the candidate answer occurs in one of the triples;
2. The name of the candidate answer occurs in a title triple;
3. The "specific words" for this answer occur in one triple at least. "Specific words" means here words which have the higher difference between their frequency in the 100 snippets corresponding to the entity and their frequency in a "real-world corpora" (here 10% of ClueWeb09);
4. The target type occurs in one of the triples.

If an URI satisfying all the constraints and not already associated for an other named entity was found, we associated it to the candidate answer with a confidence score of 1. If we did not find such URI we released the last constraint (the 4th one) and searched again with a confidence score

[4] http://mallet.cs.umass.edu/api/cc/mallet/cluster/KMeans.html

of 0.9. We repeated this process until an URI was found or until all constraints were released in which case the candidate answer was discarded.

The confidence score of an URI for a named entity E is formulated as :

$$URI(E) = 0.6 + 0.1 * nc \tag{11}$$

where nc is the number of constraints used to find the URI.

3.6 Runs

We have some criteria to estimate the relevance of a named entity and we can propose different ways to use combine them to rank candidate answers and URIs.

3.6.1 LIAcwb and LIAwb

The LIAcwb run is the only one which used the ClueWeb09 to find relevant web pages at the beginning of the process (all the others were using Bing). Only these two runs used a small subset of the criteria presented and are our baselines. A score is associated to each candidate answer E :

$$\begin{aligned} Score(E) = \sum_{P_i \in P} Compacity(E, P_i) \\ *SD(E) * URI(E) \end{aligned} \tag{12}$$

where P is the set of 500 passages retrieved with Indri. We choose to combine all the scores by the mean of multiplications because it probably was the easiest and better way for doing this (all the scores have strictly positive values). If we made a linear combination of them, we should have to find a way to normalize all these scores and computed a weight for each of them.

3.6.2 LIAwc

For this run we used all the criteria except the ones presented in 3.2 (how the distribution between snippets of the candidate answer and the ones of the target type or the examples are close). That gives for each candidate named entity :

$$\begin{aligned} Score(E) &= \sum_{P_i \in P} Compacity(E, P_i) \\ &* \; SD(E) * HLT(E) * HLN(E) \\ &* \; CS(E) * URI(E) \end{aligned} \tag{13}$$

3.6.3 LIAwd

For this run we do not used the cluster approach but instead used what we excluded from the previous run (the measure of membership of the answer to the DBPedia target type and the similarity with examples).

$$\begin{aligned} Score(E) &= \sum_{P_i \in P} Compacity(E, P_i) \\ &* \sum_{Ex_i \in EX} Similarity(E, Ex_i) \\ &* \; Membership(E, type) \\ &* \; SD(E) * HLT(E) * HLN(E) \\ &* \; URI(E) \end{aligned} \tag{14}$$

where EX is the set of examples and $type$ is the DBPedia target type.

4 Results

As the official results are not released yet, it's a difficult task to analyze the performances of our methods. Anyway, we analyze a bunch of topics by ourselves in order to pointed out the main characteristics and difficulties that our methods could have. Of course the following results and analysis haves to be considered cautiously, because the observed phenomena may not be representatives.

The Table 2 shows some precision measures for two topics [5] (22 and 51) for each run. With this only two topics this is difficult to know if using the Web as resource to find candidate answers (instead of the ClueWeb09 collection) is interesting or not. For the topic 22 using the Web seems to bring noise but for the topic 51 it appears that the coverage of the ClueWeb09 is probably not important enough and using the Web is useful. An alternative of our approach (only one resource for a given run) could be to use both or to try to determine for each topic which one have to be used (i.e. does the ClueWeb09 cover this topic?). Moreover, information on the Web change with time and for some topics the good answers are not the same now that there was at the time of the ClueWeb09 was crawled. For the topic 24 for instance, we looked for members of the "Jazz at Lincoln Center Orchestra" and the examples given in the topic are not correct anymore (and don't appear on the official Web pages) that make them (almost) useless. So, for topics for which answers could change fast it is best to use the ClueWeb09.

Table 1 and 3 show the top ten results for three topics (but not necessary for all the runs). They show that the first selection of candidate answer

[5] Official topics are available at : http://trec.nist.gov/data/entity/11/11.entity_elc_topics

LIAcwb	**LIAwb**	**LIAwc**	**LIAwd**
Hepatis	Maryland	Maryland	***National Human Genome Research Institute***
Javascript	Rockville Pike	Rockville Pike	National Diabetes Education Program
Public Health Service	Bethesda	Bethesda	***National Eye Institute***
Neuroscience Research	Rockville Pike Bethesda	Rockville Pike Bethesda	***National Cancer Institute***
Consensus Development Panel	**NIMH**	**NIMH**	**NIDA**
Wikipedia	NIH Institutes	NIH Institutes	NIH Institutes
Bethesda	Digestive	Digestive	NIH Office
National Cancer Institute	**NLM**	**NLM**	**NNCAM**
National Institutes Health	Musculoskeletal	**NIA**	**NLM**
Maryland	**NNCAM**	Musculoskeletal	NIH NHLBI Labs NHLBI

Table 1: Top ten results for topic 51. Correct results are in bold and examples are also in italic.

	Topic 22			*Topic 51*		
Run	P@5	P@10	P@R	P@1	P@5	P@10
LIAcwb	**1**	**1**	**1**	0	0	0.1
LIAwb	0.8	0.7	7/12	0	0.2	0.3
LIAwc	0.8	0.6	2/3	0	0.2	0.3
LIAwd	**1**	0.7	7/12	**1**	**1**	**0.7**

Table 2: Non official precision measure for all the runs on two topics.

Topic 24	**Topic 62**
Wynton Marsalis	American Cruise Lines
New York	Royal Caribbean
Lincoln Center Jazz Orchestra	American Cruise Line
Frederick Rose Hall	Carnival Cruises Lines
New York City Ballet	Cruise Line Cruises
Marsalis	Carnival Cruise Line
Frank Stewart	Carnival Cruises
Ahmad Jamal	Direct Line Cruises
Lincoln Center Board	Cruise Line
Wynton	Destination Cruises

Table 3: Top ten results for topic 24 (run LIAcwb) and topic 62 (run LIAwd).

is a crucial point. With topic 22 we can see that the Stanford-NER tool is not adapted. Indeed, we looked for named entities of type ”Person” (in italic) but on the top ten the precision of the NER tool is only 50%. Some tools like DBPedia Spotlight [6] seems to be more adapted here. Top ten results for topic 62 show that to merge all the different spelling of one entity is really important.

The higher results for LIAwd on topic 51 than to the ones on topic 22 seem to be due to our cooccurrences measures (the correct answers are contain in tables in the homepage of the source entity for topic 51[7]) and by our similarity measures (to the target type and to the examples). This last observation is comforted by Table 1 which shows that the top ten answers are all related to NIH and non topic related answers (eg. Maryland, Rockville, etc.) didn’t appear here. However, this method make examples ranked higher (due to our similarity function).

5 Conclusion

This paper presented our work for the ELC task of the 2011 Entity track. We tried to evaluate the impact of using results that we are confident in to rank the other results. We started with what we did in 2010 for the REF task and added new criteria in order to use the information given by the examples in the topics to rank the candidate answers.

Our unofficial evaluations show that our approach seems to improve the obtained ranking but some difficulties remain to be overcome.

References

K. Balog, P. Serdyukov, and A.P. de Vries. 2011. Overview of the trec 2011 entity track. In *NIST Special Publication : TREC 2011.*

L. Bonnefoy, P. Bellot, and M. Benoit. 2010. Liaismart at trec 2010: An unsupervised web-based approach for filtering answers. In *NIST Special Publication : TREC 2010.*

L. Bonnefoy, P. Bellot, and M. Benoit. 2011. The web as a source of evidence for filtering candidate answers to natural language questions. In *10th edition of the IEEE/WIC/ACM International Conference on Web intelligence, 2011.*

S. F. Chen and J. Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *34th Annual Meeting of the ACL, Vol. 13 (1996), pp. 310-318.*

L. Gillard, L. Sitbon, E. Blaudez, P. Bellot, and M. El-Beze. 2006. Relevance measures for question answering, the lia at qa@clef-2006. In *Lecture Notes in Computer Science,4730/2007, Evaluation of Multilingual and Multi-modal Information Retrieval , p. 440 449, 2007.*

M. Sahlgren. 2008. The distributional hypothesis. In *Special issue of the Italian Journal of Linguistics, Vol. 20 (2008).*

[6] http://spotlight.dbpedia.org/demo/index.html

[7] http://www.nih.gov/icd/