Cross-Document Cross-Lingual Coreference Retrieval

Elif Aktolga, Marc-Allen Cartright, and James Allan
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts Amherst
Amherst, Massachusetts
{elif, irmarc, allan}@cs.umass.edu

ABSTRACT

In this work, we address *coreference retrieval*, which involves identifying aliases that are distinct references to an entity. We begin with a known alias and discover unknown aliases that refer to the same entity. We use Entity Language Models to capture the contextual language around the known alias, which aids in finding new aliases. We also show that modeling the significant dates of the known aliases improves alias discovery performance.

Categories and Subject Descriptors: H3.3 [Information Storage and Retrieval]: Information Search and Retrieval - Retrieval Models

General Terms: Algorithms, Experimentation, Languages

Keywords: Coreference Analysis, Information Extraction, Information Retrieval

1. INTRODUCTION

Given a query which is a person's name, our task is to find documents with other unique references to that person. We call that task coreference retrieval. While researchers typically extract individual entities to build coreference chains, we explore this problem within an *information retrieval* framework by evaluating it as a document ranking problem. We employ entity language models (ELMs), introduced by Raghavan et al. [1]. Additionally, we use the temporal information of initially retrieved documents to from another model to aid in alias discovery. We retrieve more relevant documents than through baseline methods such as exact phrase match and bag-of-words retrieval.

We define as relevant any document that contains an alias of the person's name but not the query name. We treat an entity as an equivalence class of aliases (e.g. {*Sean Combs, Puff Daddy, P. Diddy*}). Since we are dealing with multilingual, machine-translated corpora, possible errors introduced by misspellings, mistranslations, and nicknames (pseudonyms) complicate finding new aliases.

This problem is similar in some ways to cross-document coreference analysis. We do not address reference disambiguation here, but it has been shown that ELMs provide some help there as well [2].

2. METHODOLOGY

Our ELM-based methods can be compared with standard bag-of-words retrieval (*query likelihood*, BOW), except an ELM is a probability distribution of the language surrounding all references to a particular entity [1]. Figure 1 shows the entity $E =$ 'Marge Andrews' with surrounding context words highlighted in bold italics.

> ...before trickling into the food court or the mall proper, ***as retailers started raising their security gates at 8 a.m.*** Marge Andrews ***said there was a very different feeling in the mall*** Saturday compared to her regular walks ...

Figure 1: An example reference instance

The terms in the entity model of E (ELM_E) are weighted by their maximum likelihood estimate, from which the highest-weighted j terms (i.e. t_1 is the highest weighted term, etc.) are used for calculating the score for a document D. We use Dirichlet smoothing (μ=2500) to avoid zero probabilities:

$$P(ELM_E|D) = \prod_{i=1}^{j} P(t_i|D)$$

The final score for D is an interpolation of the original query likelihood and the ELM likelihood:

$$Score_{ELM}(D) = I_D((\lambda\ P(E|D) + (1-\lambda)\ P(ELM_E|D))$$

where the indicator variable I_D is 0 if E is contained in D and 1 otherwise. This guarantees that any documents containing the original query phrase are placed at the bottom of the total ranking; they are non-relevant by construction.

We considered the soundex algorithm as a reasonable baseline; however since it is biased towards Latin-based languages, it performed poorly with our non-English collections.

As a second approach, we form a model based on the publication dates of the initial documents containing E. This way we capture documents reporting on the same entity around the same time. The significance of time is modeled in a hierarchical fashion: the closer a document's date is to that of the reference instances, the higher its weighting is

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.
ACM 978-1-59593-991-3/08/10.

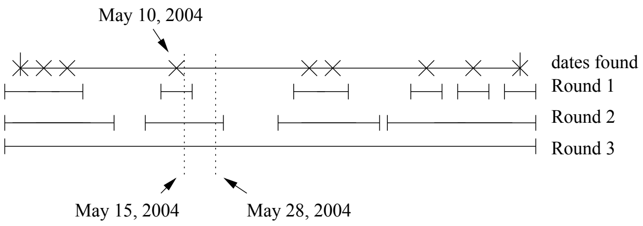

Figure 2: An example of generating date ranges.

in the model. We choose an expansion radius r for increasing the size of the date ranges. We iteratively extend the inital date ranges by (number of iteration $* r$) days, until we cover a range of (first date - r) to (last date + r) as a single range. Overlapping date ranges are merged. Figure 2 shows an example with May 10, 2004 being an initial date. A time model (TM) is thus the set of all ranges covering that period.

The TM score for a particular document D is calculated as follows. Let $\delta \in TM$ be a date range in TM:

$$Score_{TM}(D) = \frac{\|\{\delta \in TM \mid I_{TM}(D,\delta) = 1\}\|}{||TM||}$$

where the indicator function $I_{TM}(D,\delta)$ is 1 if the publish date of D is contained in δ, and 0 otherwise.

In our experiments we utilize TM by interpolating it with the ELM score:

$$Score_{TELM}(D) = \gamma\ Score_{TM}(D) + (1-\gamma)\ Score_{ELM}(D)$$

3. EXPERIMENTS & RESULTS

We use the English (93143 documents), translated Arabic (101511 documents), and translated Mandarin (57721 documents) newswire collections from the GALE version 3.12 project as our corpora. We ran all our experiments using the Indri search engine retrieving the top 1000 documents.

Table 1: Results of the baseline BOW and our approaches (ELM and TELM) for the test queries.

	BOW	*ELM*	*TELM*
P@5	0.3000	0.4083	0.3417
P@10	0.2576	0.3562	0.3177
P@20	0.2299	0.3172	0.2792
MAP	0.1784	0.1951	0.1829
NDCG@10	0.2801	0.3828	0.3331
NDCG@1000	0.3043	0.3625	0.3026

Table 1 shows the results for the test queries. The BOW baseline performs worse at all ranks. ELM and TELM perform better by retrieving documents containing pseudonyms. This is the hardest class of alias transformations, since it cannot be located through the query string itself. Success here depends solely on information surrounding the query, which we capture by means of ELMs.

The gains of our ELMs-based approach are not as pronounced as we expected. Failure analysis revealed that separate contexts are built around entities between languages. For example 'Saddam Hussein' has different contexts in English and Arabic newswire corpora. Hence, locating new aliases becomes difficult if we cannot rely on the contexts. Initial experiments on using single language corpora for the ELM construction did not improve performance on the training data.

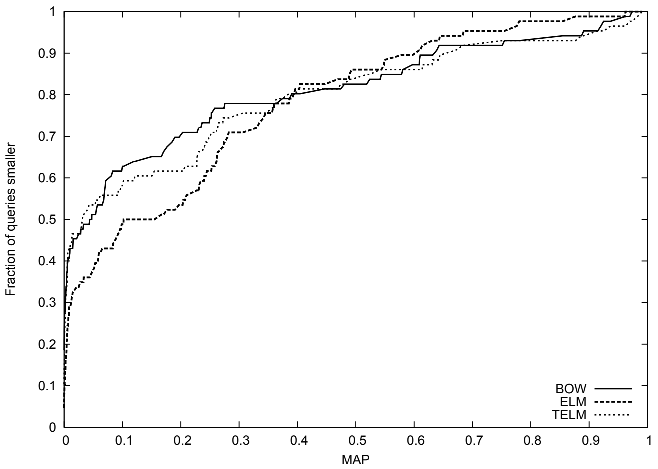

Figure 3: A graph showing the cumulative distribution of MAP with MAP values of the test queries on the x-axis and the percentage of queries with smaller MAPs on the y-axis. ELM has the smallest number of queries which have $MAP < 0.4$.

We were also surprised to find that while the TELM performed better than the BOW model, it did not perform as well as the simpler ELM. Although a full failure analysis is still pending, we conjecture that the current hierarchical model of time introduces too much noise into the scoring formula.

Figure 3 shows the fraction of queries having a MAP below a certain threshold for all the approaches. ELM has the least number of queries with low MAPs (smaller than 0.4), whereas BOW has the most number of such queries. This shows that employing ELMs for alias retrieval reduces the number of poorly performing queries overall.

For future work, we want to explore this task with passage retrieval and eventually named-entity retrieval. Ultimately, we want to arrive at a ranked list of alias names for a given alias query.

4. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by DARPA under contract #HR0011-06-C-0023. We would also like to thank our partners in the GALE project for providing machine translation and information extraction expertise. Any opinions, findings or conclusions expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

5. REFERENCES

[1] H. Raghavan, J. Allan, and A. McCallum. An exploration of entity models, collective classification and relation description. In *Proceedings of the Second International Workshop on Link Analysis and Group*, pages 1–10, 2004.

[2] C. Gooi and J. Allan. Cross-document coreference on a large scale corpus. In *Proceedings of HLT 2004 conference*, pages 9–16, 2004.