Entity based Translation Language Model

Amit Singh
IBM Research
Bangalore, India
amising3@in.ibm.com

ABSTRACT

Bridging the lexical gap between the user's question and the question-answer pairs in Q&A archives has been a major challenge for Q&A retrieval. State-of-the-art approaches address this issue by implicitly expanding the queries with additional words using statistical translation models. In this work we extend the lexical word based translation model to incorporate semantic concepts. We explore strategies to learn the translation probabilities between words and the concepts using the Q&A archives and Wikipedia. Experiments conducted on a large scale real data from Yahoo Answers! show that the proposed techniques are promising and need further investigation.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Retrieval models, Query formulation

General Terms

Algorithms, Experimentation

Keywords

question-answering, question retrieval, entity, language model

1. INTRODUCTION

Researchers have proposed the use of translation models [1, 2, 7] to address the lexical gap for the Q&A retrieval task. These models learn the translation probabilities between words using parallel mono-lingual corpora created from the Q&A pairs. While useful, effectiveness of these models is highly dependant on the availability of quality corpus [4]. Also these models only capture shallow semantics between words via the co-occurrence statistics, while some of the more explicit relationships between words and entities is freely available externally. In this paper, we propose Entity based Translation Language Model (ETLM), to accommodate this semantic information associated with entities. In this work we extend lexical word based translation model to allow for entity based query expansion. The guiding hypothesis being, entity based representation [6] reduces ambiguity of the users question and provides for a more semantically accurate expansion if relationship between entities and words can be estimated reliably. We also explore simple ways to estimate translation probabilities between entities and words and report our findings on a large scale real data from Yahoo Answers comprising ∼5 million Q&A pairs.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

2. PROBLEM DEFINITION & APPROACH

Let $C = d_1, d_2, ..., d_n$ denote the Q&A collection; d_i refers to the i-th Q&A data consisting of a question q_i and its answer a_i. Given the user question q_{user}, the task of Q&A retrieval is to rank d_i according to $score(q_{user}, d_i)$. Approach to compute this score in the ETLM framework is outlined as follows. 1) Annotate references of Wikipedia entities in Q&A archives. 2) Compute translation probabilities between entities and words using Q&A pairs and Wikipedia. 3) At runtime, annotate the user query (q_{user}) with Wikipedia entities to create q and then rank d_i using $score(q, d_i)$ calculated as per ETLM model described below.

ETLM Model: Let the annotated query q (and d_i) be composed of sequence of token spans (T_q). Each token span (t_q) corresponds to sequence of contiguous words occurring in the running text. These t_q's can correspond to entity mentions, phrases or words. Let e_q denote the tokens spans that are annotated and ne_q that are not ($T_q = e_q \bigcup ne_q$). For example, in the query , $\underbrace{\texttt{What}}_{ne_q}\ \underbrace{\texttt{is}}_{ne_q}\ \underbrace{\texttt{a}}_{ne_q}\ \underbrace{\texttt{Quadratic Formula}}_{e_q}\texttt{?}$, token span `Quadratic Formula` is linked to Wikipedia entity corresponding to Quadratic Equation[1] while all other token spans are marked as ne_q . For the sake of simplicity, in this work we do not identify phrases i.e. ne_q is always of unit word length. In the ETLM framework, the similarity between a query q and a document d within a collection C is given by the probability

$$score(q,d) \sim P(q|d) = \prod_{\substack{t_q \in q \\ t_q = e_q \bigcup \mathrm{ne}_q}} P(t_q|d) \quad (1)$$

$$P(t_q|d) = (1-\lambda) \sum_{t_d \in d} T(t_q|t_d) P_{ml}(t_d|d) + \lambda P_{ml}(t_q|C) \quad (2)$$

$T(t_q|t_d)$ denotes the probability that a token span t_q is the translation of token span t_d. This induces the desired query expansion effect. It is evident that if $|e_q| = 0$ then ETLM reduces to Translation Model [2]. The key task is to estimate $P_{ml}(t_q|C)$, $T(t_q|t_d)$ and $P_{ml}(t_d|d)$; $t_q \in e_q \bigcup ne_q$ and $t_d \in e_d \bigcup ne_d$

Estimating from parallel corpus ETLM^{qa}: Following [7] we pool the question and answers to create the parallel corpus. We constructed a parallel corpus C_0 as $(q_1, a_1), \ldots, (q_n, a_n) \bigcup (a_1, q_1), \ldots, (a_n, q_n) = C_0$. To handle entities e, we introduce special id's in the ne space so that $T(t_q|t_d)$ is

[1] http://en.wikipedia.org/wiki/Quadratic_equation

learnt w/o any modification to the corresponding translation algorithm. Remaining model components are calculated using Equation 3 and 4. Here d refers to question part of the Q&A pair.

$$P_{ml}(t_q|C) = \frac{tf_{t_q,C}}{\sum_{t' \in C} tf_{t',C}} \quad (3)$$

$$P_{ml}(t_q|d) = \frac{tf_{t_q,d}}{\sum_{t' \in d} tf_{t',d}} \quad (4)$$

Estimating from Wikipedia ETLM^{wiki}: Number of symmetric measures have been proposed to measure semantic relationships between entities and words using Wikipedia. For our problem we need an asymmetric measure. We use co-citation information in Wikipedia to detect relatedness between entities ($T(e|e')$) and co-occurrence counts to estimate $T(ne|ne')$ as follows:

$$T(e|e') = \frac{co(e,e')}{\sum_{e''} co(ne'',ne')} \quad (5)$$

$$T(ne|ne') = \frac{df(ne,ne')}{\sum_{ne''} df(ne'',ne')} \quad (6)$$

$$T(ne|e) = \frac{tf_{ne,d(e)}}{|d(e)|} \quad (7)$$

$$T(e|ne) = \frac{tf_{ne,d(e)}}{\sum_{e' \in E} tf_{ne,d(e')}} \quad (8)$$

Here $d(e)$ represents the page corresponding to entity e. $df(ne,ne')$ is the number of Wikipedia page containing both ne and ne'. $tf_{t,d(e)}$ is the frequency of t in $d(e)$; $co(e,e')$ indicates number of entities in Wikipedia that have a hyperlink to both e and e' To make sure self translation probability is not underestimated i.e. $T(t|t) \geq T(t'|t)$ always holds true, we introduce new parameter γ in Equation 5 and 6 as $T(t|t') = \gamma + (1-\gamma)T(t|t')$; $\gamma = 0$ when $t \neq t'$ and $\gamma > 0.5$ otherwise. $P_{ml}(t_q|C)$ and $P_{ml}(t_q|d)$ are estimated as per Equation 3 and 4 respectively.

Wikipedia Entity Annotator (EA): We used LOCAL+ PRIOR [3] configuration to annotate $d_i \in C$ and q_{user}. Parameter $\rho_{na} \geq 0$ was the knob used to control the back-off strategy for annotation. ρ_{na} was tuned over 83 Q&A pairs, amounting to 412 human tagged annotations. $\rho_{na} = 6.4$ corresponding to Precision=0.91 Recall=0.54 was used for evaluation presented in Section 3.

3. EVALUATION

We used a dataset consisting of ∼5 million questions and corresponding best answers from Yahoo! Answers spanning all the leaf level categories. In our retrieval experiments we used 237 queries (average length=5.1 words).For each of the queries, we pooled the top 25 Q&A pairs from retrieval results generated by varying the retrieval algorithms and the search field. Relevance judgments were marked by the human annotators without disclosing the identity of method used for retrieval. Semantic similarity of the question to the user query was the only criteria used for marking relevance of a Q&A pair. Over all we had collected more than 1950 relevance Q&A pairs corresponding to 237 queries. To evaluate performance of our retrieval system we used standard IR metrics (MAP, MRR, R-Precision, Precision@5) measured on recall base of 20. We evaluated four baseline retrieval models VSM [8], OKAPI BM25 [5] TLM [2] and TransLM(using answers) [7]. For our experiments we used a set of 35 queries to select the model parameters: λ and γ. Final values of ETLM parameters used in our experiments were $\lambda = 0.78$ and $\gamma = 0.62$.

Method	MAP	MRR	R-Precision	Precision@5
VSM	0.251	0.452	0.242	0.23
OKAPI BM25	0.352	0.551	0.32	0.29
TLM	0.381	0.571	0.371	0.322
TransLM	0.394	0.575	0.378	0.331
ETLM^{qa}	0.418*	0.585	0.391*	0.357*
ETLM^{wiki}	0.441*†	0.613*†	0.417*†	0.371*†

Table 1: Comparisons with four baseline retrieval models. $*$ and † indicate a statistically significant improvement over the TransLM and ETLM^{qa} respectively using paired t-test with p-value < 0.05.

Result Analysis Performance of all the translation based models is better than VSM and OKAPI thereby confirming the importance of addressing the lexical gap. Using high confidence annotations for query expansion in ETLM, leads to an improved performance as compared to the all the four baseline methods that do not consider this signal. This is validated by the fact that ETLM^{qa} and ETLM^{wiki} can achieve statistically significant improvements in terms of most of the measures. The reason for this improvement is the context sensitive computation of $T(t|t')$ leading to reduced spurious expansions and improved top expansions, this is made possible because of entity disambiguation. This computation in baseline is done on a word by word basis without exploiting contextual information. ETLM^{qa} performs worse than ETLM^{wiki} because of the sparsity of the annotations (high value of ρ_{na}), that fail to capture rich relationship made easily available in Wikipedia.

4. REFERENCES

[1] A. Berger and J. Lafferty. Information retrieval as statistical translation. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 222–229, New York, NY, USA, 1999. ACM.

[2] J. Jeon, W. B. Croft, and J. H. Lee. Finding similar questions in large question and answer archives. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 84–90, New York, NY, USA, 2005. ACM.

[3] S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti. Collective annotation of wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–466. ACM, 2009.

[4] J.-T. Lee, S.-B. Kim, Y.-I. Song, and H.-C. Rim. Bridging lexical gaps between queries and questions on large online q&a collections with compact translation models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 410–418, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[5] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. pages 109–126, 1996.

[6] A. Singh, S. Kulkarni, S. Banerjee, G. Ramakrishnan, and S. Chakrabarti. Curating and searching the annotated web. In *In SIGKDD Conference, 2009. System demonstration.*

[7] X. Xue, J. Jeon, and W. B. Croft. Retrieval models for question and answer archives. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 475–482, New York, NY, USA, 2008. ACM.

[8] J. Zobel and A. Moffat. Inverted files for text search engines. *ACM Comput. Surv.*, 38, July 2006.