Entity based Q&A retrieval

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 the 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. While useful, the effectiveness of these models is highly dependant on the availability of quality corpus in the absence of which they are troubled by noise issues. Moreover these models perform word based expansion in a context agnostic manner resulting in translation that might be mixed and fairly general. This results in degraded retrieval performance. In this work we address the above issues by extending the lexical word based translation model to incorporate semantic concepts (entities). We explore strategies to learn the translation probabilities between words and the concepts using the Q&A archives and a popular entity catalog. Experiments conducted on a large scale real data show that the proposed techniques are promising.

1 Introduction

Over the past few years community-based question answering (CQA) portals like Naver, Yahoo! Answers, Baidu Zhidao and WikiAnswers have attracted great attention from both academia and industry (Adamic et al., 2008; Singh and Visweswariah, 2011). These portals foster collaborative creation of content by allowing the users to both submit questions to be answered and answer questions asked by other users. These portals aim to provide highly focused access to this information by directly returning pertinent question and answer (Q&A) pairs to the users questions, instead of a long list of ranked URLs. This is in noted contrast to the usual search paradigm, where the question is used to search the database of potential answers, in this case the question is used to search the database of previous questions, which in turn are associated with answers. This involves addressing the word mismatch problem between the users question and the question-answer pairs in the archive. This is the major challenge for Q&A retrieval.

Researchers have proposed the use of translation models (Berger and Lafferty, 1999; Jeon et al., 2005; Xue et al., 2008) to solve this problem. As a principled approach to capturing semantic word relations, statistical translation language models are built by using the IBM model 1 (Brown et al., 1993) and have been shown to outperform traditional document language models on Q&A retrieval task. The basic idea is to estimate the likelihood of translating a document[1] to a query by exploiting the dependencies that exists between query words and document words. For example the document containing the word `Wheezing` may well answer the question containing the term `Asthma`. They learn the these dependencies (encoded as translation probabilities) between words using parallel mono-lingual corpora created from the Q&A pairs. While useful, the effectiveness of these models is highly dependant on the availability of quality corpus (Lee et al.,

[1] we will use (Q&A, document), (word, term) and (user query, question) interchangeably

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1266–1277, Jeju Island, Korea, 12–14 July 2012.

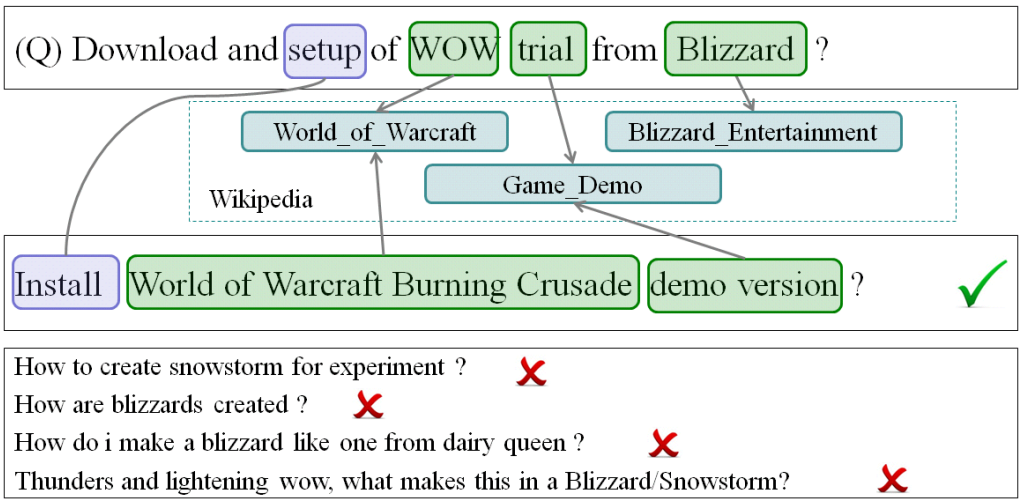

Figure 1: Need for entity based expansions

2008). 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. Being context agnostic (Zhou et al., 2007) is another very common criticism hailed on translation models as it results in noisy and generic translations.

Example shown in Figure 1 captures these problems. Specifically, the word `Blizzard` can refer to an American game development company that develops World of Warcraft game or it could refer to a severe snowstorm. Expanding query without taking the gaming context established by the word `WOW` (acronym for World of Warcraft) into account would lead to topic drift. Also it would be difficult to learn relationships between `World of Warcraft Burning Crusade` and `Blizzard` from the Q&A corpus alone due to the sparsity of co-occurance counts as these can be expressed in several lexical forms, some of which are multi word phrases.

In this paper we argue that solution to all the above problems lies in a unified model in which entities are a primary citizen. The guiding hypothesis being, an entity based representation provides a less ambiguous representation of the users question and provides for a more semantically accurate expansion if the relationship between entities and words can be estimated more reliably. Our main contributions are

1. We propose Entity based Translation Language Model (ETLM) for Q&A retrieval that accommodates semantic information associated between entities and words. Being closely related to the general source-channel framework (Berger and Lafferty, 1999), the model enjoys its benefits, while mitigating some of its shortcomings. Specifically it provides for context aware expansions of the query by exploiting entity annotations on both, the document and the query side. Entity annotations also provide a means to handle the "many-to-one" (Moore, 2004) translation limitation in the IBM model, due to which each word in the target document can be generated by at most one word in the question[2]. For the same reasons, it also alleviates another related limitation by enabling translation between contiguous words across the query and documents (Moore, 2004).

2. We learn relationships between entities and terms by proposing new ways of organizing monolingual parallel corpus and simultaneously leveraging external resources like Wikipedia from which one can derive these relationships reliably. This helps alleviate the noise problem associated with learning translation models on Q&A archive described above. An important point to note is that, our technique has merits independent to the choice of the entity catalog. In this work we use

[2]entity mentions can be of more than unit word length

D	original Q&A collection
E	set of all entities in catalog
$d(e)$	description of entity e
C	D annotated with $e \in E$
q_{user}	users question
q	q_{user} annotated with $e \in E$
t	token span
t_q, t_d	token span in q and d
V	word vocabulary

Table 1: Notation.

Wikipedia, as it is a popular choice due to its large and ever expanding coverage and its ability to keep up with world events on a timely basis.

3. We provide detailed evaluation of impact of modelling assumptions and model components on retrieval performance on a large scale real data from Yahoo Answers comprising ∼5 million Q&A pairs.

Rest of the paper is organized as follows: In the next section, we define ETLM and outline its details. This is followed by Section 3 which gives the details of entity annotators and its performance. Section 4 describes our experiments on the retrieval method used Q&A retrieval. In Section 5 we compare and contrast related literature. Finally, we conclude in Section 6.

2 Our Approach

Problem Definition: Let $D = d_1, d_2, ..., d_n$ denote the Q&A collection. Here 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 $score(q_{user}, d_i)$. Figure 2 outlines the approach to compute this score in the ETLM framework.

Offline processing: Using the entity catalog E, we learn the entity annotation models $\text{EA}_{offline}$ and EA_{online} for annotation of entities in the Q&A corpus and the query respectively. Refer Section 3 for details. For each $d_i \in D$, we then annotate references to entities in Wikipedia using $\text{EA}_{offline}$ resulting in annotated Q&A corpus C. We then compute relationships between entities and words using C and E. These relationships are used to learn our ETLM model.

Online processing: At runtime, annotate the user query q_{user} with entities using $\text{EA}_{online}$to create an enriched question q. Issue this query over the annotated corpus C and rank the candidates as per the ETLM model described below.

2.1 ETLM Model

Let the annotated query q (and similarly annotated Q&A pair d) be composed of sequence of token spans T_q (and T_d). Each token span t_q (similarly t_d) 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 ($T_q = e_q \cup ne_q$). 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 an entity corresponding to Quadratic Equation[3], 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[4]. In the ETLM framework, the similarity between a query q and a document d within a collection C is given by the probability

$$
\begin{aligned}
score(q,d) \quad &\sim \quad P(q|d) = \prod_{\substack{t_q \in q \\ t_q = e_q \cup \text{ne}_q}} P(t_q|d) \\
P(t_q|d) \quad &= \quad (1-\lambda) P_{ml}(t_q|d) + \lambda P_{ml}(t_q|C) \\
P_{ml}(t_q|d) \quad &= \quad \sum_{t_d \in d} T(t_q|t_d) P_{ml}(t_d|d) \qquad (1)
\end{aligned}
$$

Intuitively this indicates a generative process for creating q from d. Ideally both q and d are "only" composed of e i.e. $\forall t_q \in q;\ t_q \in E_U$, where E_U is the universal set of entities [5] (similarly for all t_d). This is because when the document was created, each and every $t_d \in d$ had a sense attached to it. however in reality, for various reasons, set of target entities are clearly a subset of E_U (for e.g. E: set of all entities

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

[4] its not a restriction as the model is valid for ne_q consisting of more than one word.

[5] language for creating q and d

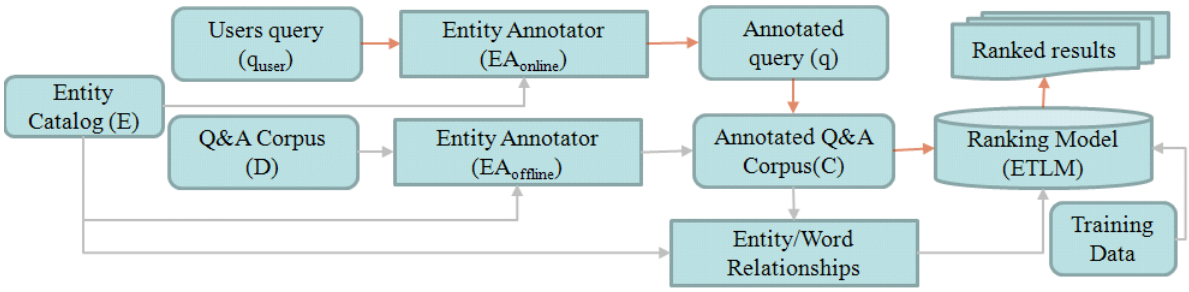

Figure 2: ETLM Architecture (gray and brown arrows indicate offline and online processes respectively

in the catalog) also not all of them may be recognized by the annotation system.

$T(t_q|t_d)$ in Equation 1 denotes the probability that a token span t_q is the translation of token span t_d. This induces the desired query expansion effect. The key task is to estimate $P_{ml}(t_q|C)$, $T(t_q|t_d)$ and $P_{ml}(t_d|d)$; $t_q \in e_q \cup ne_q$ and $t_d \in e_d \cup ne_d$

2.2 Estimating Model Parameters

We adopt 2 different approach for estimating $T(t_q|t_d)$, leading to 2 different configurations of ETLM system . As the name suggests, ETLM$_{qa}$ is estimated from Q&A data (C and D) while we leverage the entity catalog (in our case it is Wikipedia) for ETLM$_{wiki}$.

2.3 ETLM$_{qa}$: Estimate from parallel corpus

Following (Xue et al., 2008) we pool the question and answers from D to create a master parallel corpus $P = (q_1, a_1), .., (q_n, a_n) \cup (a_1, q_1), ., (a_n, q_n)$. This is used for learning $T(ne|ne')$[6]. Similarly we create P^* from C. We then derive 2 different parallel corpora from P and P^* as follows

P_{entity} We remove all non linked tokens ne from P^* thereby reducing it to parallel corpus over e. This is used for learning $T(e|e')$ i.e. translation probabilities between two entities e and e' in E.

P_{hybrid} This is hybrid of P_{entity} and P where in one part of Q&A pair consists on only ne while other consists of only e. This is used for learning $T(ne|e)$ and $T(e|ne)$.

To handle entities e, we introduce special id's in the ne space. Thus our universal token span set is given by $V \cup E$. This is done so that $T(t_q|t_d)$ is learnt from P, P_{entity} and P_{hybrid}, w/o any modification to the corresponding translation algorithm (Brown et al., 1993). Lets call this approach ETLM$_{qa'}$.

We explored another intuitive approach ETLM$_{qa}$, to learn $T(e|e')$, $T(e|ne)$, $T(ne|e)$ and $T(ne|ne')$ directly by using only P^* as our parallel corpus. We do so by redistributing the probability mass i.e. when calculating $T(e|e')$, we redistribute probability mass spread over all the ne to e given by Equation 2 and 3. Similar process is followed for $T(e|ne)$, $T(ne|e)$ and $T(ne|ne')$.

$$S(e|e') = \frac{T(e|e')}{\sum_{t \in V} T(t|e')} \quad (2)$$

$$T(e|e') = \frac{S(e|e')}{\sum_{t \in E} T(t|e')} \quad (3)$$

Remaining model components are calculated using Equation 4 and 5. Here d refers to question part of the Q&A pair.

$$P_{ml}(t_q|C) = \frac{tf_{t_q,C} + 1}{\sum_{t' \in C} tf_{t',C} + |C|} \quad (4)$$

$$P_{ml}(t_q|d) = \frac{tf_{t_q,d}}{\sum_{t' \in d} tf_{t',d}} \quad (5)$$

2.4 ETLM$_{wiki}$: Estimating from Wikipedia

Number of symmetric measures have been proposed (Medelyan et al., 2009) 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

[6] subscript of q and d has been dropped as translation probability learnt agnostic to it, due to pooling.

$T(ne|ne')$ as follows: .

$$T(e|e') = \frac{co(e,e')}{\sum_{e''} co(e'',e')} \quad (6)$$

$$T(ne|ne') = \frac{cf(ne,ne')}{\sum_{ne''} cf(ne'',ne')} \quad (7)$$

$$T(ne|e) = \frac{tf_{ne,D(e)}+1}{|D(e)|+|V|} \quad (8)$$

$$T(e|ne) = \frac{tf_{ne,D(e)}+1}{\sum_{e'\in E} tf_{ne,D(e')}+|E|} \quad (9)$$

Here $d(e)$ represents the page corresponding to entity e. $D(e)$ represents concatenation of $d(e)$ and all context of size 5 surrounding anchor text in Wikipedia that link to e. $cf(ne,ne')$ is the number context windows of fixed size containing both ne and ne' in Wikipedia. In our case, we set the window size at 10 (because this size turned out to be reasonable in our pilot experiments). $tf_{t,d(e)}$ 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'. As links from pages with a small number of outgoing links are generally considered to be more valuable than links from pages with a high outgoing link degree, we tried with weighted version of 6 where the co-citations are weighted by the outdegree of Wikipedia page corresponding to entity s that link to e and e'. Lets denote the weighted version by ETLM$_{wiki}$ and unweighted version by ETLM$_{wiki'}$. $P_{ml}(t_q|C)$ and $P_{ml}(t_q|d)$ are estimated as per Equation 4 and 5 respectively.

2.5 Self translation probability

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 as $T(t|t') = \gamma + (1-\gamma)T(t|t')$; $\gamma = 0$ when $t \neq t'$ and $\gamma > 0.5$ otherwise.

2.6 ETLM$_{combo}$: Combining ETLM$_{qa}$ and ETLM$_{wiki}$

Often, combining language models yields better results than any of the individual language models themselves. Linear interpolation is often the technique of choice in language modelling for combining models to exploit complementary features of the component models. It involves taking a weighted sum of the probabilities given by the component language models. An advantage of the linear interpolation is that it is simple and fast to calculate. If the inputs are probability estimates, also the output is a probability estimate. The mixture translation model $T_{combo}(e|e')$ over M component models is given by Equation 10.

$$T_{combo}(t|t') = \sum_{j=1}^{M} \alpha_j T_j(t|t')$$
$$t \in E \cup V; \quad \sum_{j=1}^{M} \alpha_j = 1; \quad \alpha_j \geq 0 \quad (10)$$

One can immediately notice that $T_{combo}(t|t')$ has one global weight for each of the M component models which might not be ideal. With access to large training data one could employ more powerful context dependent interpolation techniques (Liu et al., 2008). In our case we have 2 components T$_{qa}$ and T$_{wiki}$ and four classes for each ; $\alpha_{wiki}^{(e,e')}$[7], $\alpha_{wiki}^{(e,ne)}$, $\alpha_{wiki}^{(ne,ne')}$ and $\alpha_{wiki}^{(ne,e)}$), one corresponding to each class of $T(t|t')$. respectively.

3 Entity Annotation

In this section we describe our entity annotation system. Recently there has been lot of work addressing the problem of annotating text with links to Wikipedia entities (Mihalcea and Csomai, 2007; Bunescu and Pasca, 2006; Milne and Witten, 2008; Kulkarni et al., 2009; Ratinov et al., 2011; Ferragina and Scaiella, 2010). We adopt a similar approach, wherein we first find the best disambiguation (BESTDISAMBIGUATION) for a given mention and then decide to prune it (PRUNE), via the dummy mapping NA (similar to "no assignment" (Kulkarni et al., 2009)).

3.1 BESTDISAMBIGUATION

As defined earlier, $e \in E$ represent an entity corresponding to URN of a Wikipedia article. Let $E_m = \{e_{m,1}, e_{m,2}, \cdots, e_{m,|E_m|}\}$ $e_{m,i} \in E$ represent the set of possible disambiguations for a mention m (m is an index over all mentions in the corpus). Given a mention m, task is to find best disambiguation e from Wikipedia. Without loss of gener-

[7] $\alpha_{qa}^{(e,e')} = 1 - \alpha_{wiki}^{(e,e')}$

ality, we consider $e_{m,*} \in E_m$ as the correct answer. Let $\phi(m, e_{m,j})$ represent the mapping onto features between an entity mention m and the Wikipedia entity $e_{m,j}$ and $\overrightarrow{\omega}$ be the corresponding weight vector and $D(e_{m,j}) = \overrightarrow{\omega}\phi(m, e_{m,j})$ represent the disambiguation score. The task is to learn $\overrightarrow{\omega}$ such that $\underset{e_{m,j}}{\text{argmax}}\ D(e_{m,j})$ gives the best disambiguation for the mention m.

We pose this as a ranking problem and solve it using max-margin technique (Joachims, 2002; Joachims, 2006) as follows

$$\begin{aligned} &\underset{\overrightarrow{\omega}, \overrightarrow{\xi}}{\text{minimize}} \qquad \frac{1}{2}\overrightarrow{\omega} \cdot \overrightarrow{\omega} + C \sum \xi_{i,j} \\ &\text{subject to} \\ &\forall m, \forall e_{m,j} \in E_m - e_{m,*} : \\ &\overrightarrow{\omega}\phi(m, e_{m,*}) > \overrightarrow{\omega}\phi(m, e_{m,j}) + \xi_{i,j} \\ &\qquad\qquad \forall i, \forall j : \xi_{i,j} \geq 0 \end{aligned} \tag{11}$$

where $\sum \xi_{i,j}$ is the total training error that upper bounds the number of pair preferences violations. This is controlled by adjusting the parameter C. Note that Equation 11 means pairwise comparison between the correct disambiguation $e_{m,*}$ and other disambiguation candidates $e_{m,j}$ such that $j \neq$ index corresponding to *.

3.2 PRUNE

The disambiguation phase produces one candidate disambiguation per mention. To discard any unmeaningful annotations a simple strategy similar to LOCAL (Kulkarni et al., 2009) is followed where the $D(e_{m,*})$ is compared against a predefined threshold ρ_{na}, so that if $D(e_{m,*}) < \rho_{na}$ then that annotation for menton m is discarded by linking m to NA. The parameter ρ_{na} allows the algorithm to back-off when short of evidence.

3.3 FEATUREMAP $\phi(m, e_{m,j})$

Sense probability prior (SP): It represents the prior probability that a mention name s points to a specific entity in Wikipedia. For example, without any other information, mention name "tree" will more likely refer to the entity woody plant[8], rather than the less popular notion related to graphs [9].

Entity Probability prior (EP): It captures the popularity knowledge as a distribution of entities, i.e., the $EP(e_i)$ should be larger than $EP(e_j)$ if e_i is more popular than e_j. This score is independent of the mention name.

Context specific features: It captures the textual similarity between weighted word vectors corresponding to the context of the mention (window around the mention) and textual description associated with the entity (Wikipedia page).

Let EA_{online} and $\text{EA}_{offline}$ represent configurations for annotating user question and corpus respectively. For EA_{online}, user question represents the document from which context specific features are computed. For $\text{EA}_{offline}$, question and the answer(best) is concatenated to represents the document. Based on the "one sense per discourse" assumption, one additional heuristic is used in $\text{EA}_{offline}$ where, for the same Q&A pair, if same name mention is repeated multiple times across the question and the answer then one with the maximum $D(e_{m,*}) > \rho_{na}$ is annotated for all instances.

3.4 Annotation Experiments

We used 2010 version of Wikipedia as our knowledge base. It contains more than 2.5 million entities. Annotations were done by volunteers fluent in english. Volunteers were told to be as exhaustive as possible and tag all possible name mentions, even if to mark them as "NA". Inter-annotator agreement=92.1%; Kappa coefficient = 0.72. As our corpus, we collected 8.3K manual annotations spanning 1315 Q&A pairs. 2.8K of the annotations were assigned to NA. 2.1K annotations (out of 8.3K) were made in the question of which 551 were assigned to NA. We use Precision, Recall and F_1 score micro-averaged across documents as the evaluation measures. We do a linear scan of data to identify entity mentions by first tokenizing and then identifying token sequences that maximally match an entity ID in the entity name dictionary (constructed using Wikipedia anchor text, redirect pages). Figure 3 outlines the performance of $\text{EA}_{offline}$ and EA_{online}. We measured $\text{EA}_{offline}$ in 3 test data configurations; (1) $\text{EA}_{offline}$: measured over entire

[8] en.wikipedia.org/wiki/Tree

[9] en.wikipedia.org/wiki/Tree_(data_structure)

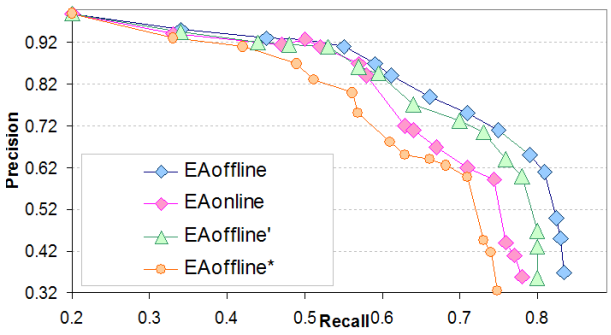

Figure 3: Precision v/s Recall

annotation set; 2) $\text{EA}_{offline'}$ is measured only on annotations made in question. this is done to compare it with EA_{online}; 3) $\text{EA}_{offline^*}$ is similar to (2), only difference is that for (2) entire Q&A pair is the context, while here only question part is the context. This is done to check if separate annotators are required for online and online phase. As seen in Figure 3, this indeed is necessary as $\text{EA}_{offline^*}$ performs worse than EA_{online}. Closer look at the feature weights revealed that in $\text{EA}_{offline}$ context specific features have much more weightage when compared to its weight in EA_{online}, on the contrary EA_{online} weighs SP significantly higher.

4 Evaluation

We now describe the empirical evaluation where we compare our techniques against the baseline techniques. We use several standard measures (R-Precision, MAP, MRR, Precision@k) in evaluation. We first describe the dataset used followed by describing an exhaustive set of results across techniques and performance measures.

4.1 Dataset

We crawled a dataset of ∼5 million questions and answers from Yahoo! Answers spanning all the leaf level categories. Tokenization and stop word removal were the only preprocessing steps performed. We have used a stoplist[10] having a vocabulary of 429 common words to remove the stopwords.

In our retrieval experiments we used 339 queries (average length 5.6 words). We employed pooling technique used in the TREC conference series. 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 human annotators without disclosing the identity of method used for retrieval. The annotators were asked to label candidate as "relevant" or "irrelevant" based on semantic similarity with the query. Answer quality/correctness was not a criteria. In case of disagreement between two volunteers, authors made the final judgment. Inter-annotator agreement was 87.9% and Kappa coefficient = 0.68. Over all we had collected more than 12K relevance judgements corresponding to these queries, of which >2.3K were marked as relevant.

[10] http://truereader.com/manuals/onix/stopwords1.html

4.2 Baselines

To evaluate the effectiveness of our models we compared them against the following baselines

Traditional models: VSM (Zobel and Moffat, 2006) and OKAPI BM25 (Robertson et al., 1996) (k1, b, and k3 are parameters that are set to 1.2, 0.75 and ∞ respectively).
Translation based language models: TLM (Jeon et al., 2005), TransLM (using answers) (Xue et al., 2008) and CTM (Lee et al., 2008).

For our experiments we used a set of 50 queries to select the model parameters. Translation based language model use 2 parameters; smoothing parameter λ in the Language Model and β to control the self-translation impact in the TransLM. Final values of parameters used in our experiments were $\lambda = 0.2$ (Zhai and Lafferty, 2004) and $\beta = 0.75$ (Xue et al., 2008). For CTM, we used *tf-idf* based weighing scheme (Lee et al., 2008) to remove words from the (Q‖A) corpus P. Word elimination threshold of 20% was selected based on the above 50 queries. Final values of ETLM parameters used in our experiments were $\lambda = 0.18$ and $\gamma = 0.65$.

4.3 Result Analysis

Table 2 presents the performance of the various techniques. Under each measure, we highlight the best performing technique. 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

	MAP	%chg	MRR	%chg	R-Prec	%chg	Prec@5	%chg	Prec@10	%chg
VSM	0.221		0.421		0.21		0.202		0.15	
OKAPI	0.298		0.532		0.271		0.264		0.214	
TLM	0.337		0.583		0.318		0.297		0.239	
TransLM	0.352		0.612		0.347		0.332		0.261	
CTM	0.361		0.641		0.351		0.341		0.279	
ETLM_{qa}	$0.390^{\dagger}$	8.03	$0.699^{\dagger}$	9.05	$0.379^{\dagger}$	7.98	$0.367^{\dagger}$	7.62	$0.302^{\dagger}$	8.24
ETLM_{wiki}	$0.413^{\dagger}$	14.40	$0.719^{\dagger}$	12.17	$0.399^{\dagger}$	13.68	$0.391^{\dagger}$	14.66	$0.323^{\dagger}$	15.77
ETLM_{combo}	$\mathbf{0.427}^{\dagger}$	18.28	$\mathbf{0.726}^{\dagger}$	13.26	$\mathbf{0.413}^{\dagger}$	17.66	$\mathbf{0.407}^{\dagger}$	19.35	$\mathbf{0.331}^{\dagger}$	18.64

Table 2: Comparisons of retrieval models. † indicate a statistically significant improvement over the CTM using paired t-test with p-value < 0.05. %chg indicates change over CTM as it is the most competitive baseline

query expansion in ETLM, leads to an improved performance as compared to the all the 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 all 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 baselines happens on word by word basis without exploiting contextual information. ETLM_{qa} performs worse than ETLM_{wiki}. On close inspection of failure cases and translation probability tables we found that $T(e|e')$ for ETLM_{qa} was much worse than ETLM_{wiki}. This is because for getting good estimates of $T(e|e')$, we need enough instances where both e and e' need to be correctly annotated in the same Q&A pair. Failure in this leads to sparse counts thereby reducing the gap in $T(e|e')$ scores for related and unrelated e. Figure 4 shows the impact of choices made for learning the translation probabilities $T(t|t')$. We found that ETLM_{wiki} performs slightly better than $\text{ETLM}_{wiki'}$, indicating the utility of weighted co-citation measure for computing $T(e|e')$. We believe that embedding other measures that are better in capturing relationships from Wikipedia, should improve the performance. Similarly ETLM_{qa} also performs better than $\text{ETLM}_{qa'}$. This is because for creating P_{entity} all ne are removed. This leads to count sparsity problem discussed above, but slightly worse. Due to absence of ne, in $\text{ETLM}_{qa'}$ e' in d are thought be being generated "only" from e in q. On the contrary in ETLM_{qa}, e' had an option of mapping to ne in q. An interesting observation is that while the performances of different configurations vary, all of them perform better than CTM which is the best baseline.

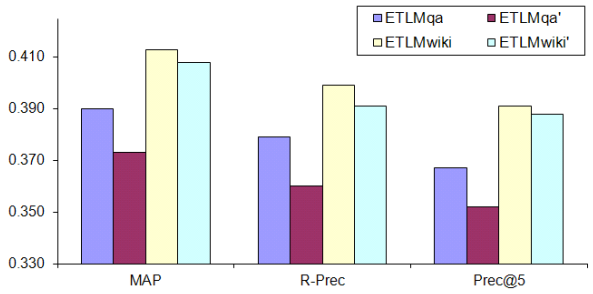

Figure 4: Performance of ETLM configurations

4.4 Impact of Annotations on retrieval

Since entities are central in our model, impact of entity annotation on q_{user} is one of the most important aspect to be studied. Figure 5 shows the plot of retrieval measures obtained by varying ρ_{na} in EA_{online}. CTM is shown by horizontal lines. As explained in Section 3, value of ρ_{na} is inversely proportional to aggressiveness of annotation. Which implies for high values, EA_{online} will annotate only those mentions in query that its highly confident about. Beyond a value no annotations are made.

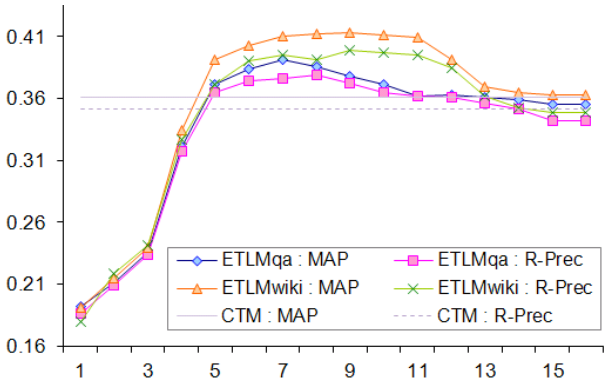

Figure 5: Impact of query annotation on retrieval. x-axis represents value of ρ_{na} used to control annotation

This is represented on the extreme right in Figure 5. One important observation is that, even with no annotations made in query, performance of ETLM$_{qa}$ and ETLM$_{wiki}$ is competitive to CTM. This is evidence for addressing the noise related issue for which CTM is designed. For large range of values, all ETLM configurations are above CTM. As we decrease ρ_{na} performance increases smoothly and then after a certain point ($\rho_{na} = 5$) is starts decreasing.

4.5 ETLM$_{combo}$

We believe that $T(t|t')$ learnt from one source would encode word association characteristics which might not be exactly the same across sources.ETLM$_{combo}$ tries to address this by combining the two models. Values for mixing parameters are : $\alpha_{wiki}^{(e,e')} = 0.95$[11], $\alpha_{wiki}^{(e,ne)} = 0.75$, $\alpha_{wiki}^{(ne,ne')} = 0.7$ and $\alpha_{wiki}^{(ne,e)} = 0.75$). The interpolation weights were obtained by optimizing the retrieval performance by doing a using grid search over the parameter space. Same 50 queries were used for tuning. As seen entity relationships obtained from Wikipedia are far superior to one from Q&A corpus. As seen in Table 2 combining the two models improves the performance.

5 Related Work

Recently Q&A retrieval has been garnering lot of attention. Translation model (TLM) (Jeon et al., 2005) has been extensively employed in question search and has been shown to outperform the traditional IR methods significantly (VSM, BM25, LM). Existing work can be broadly grouped under the following topics:

(a) *Improved training of translation models by exploiting answer content/inter-word co-occurrence relations and restriction to reliable parallel corpora:* Translation-based language model (TRLM) (Xue et al., 2008) improved stability of TLM by providing better probability estimates and also exploited answers for question retrieval. It further improved the retrieval results and obtained the state-of-the-art performance. Another line of work on translation models focused on providing suitable parallel data to learn the translation probabilities. Compact translation models (CTM) (Lee et al., 2008) tried to further improve the translation probabilities based on question-answer pairs by selecting the most important terms to build compact translation models. We show that such special-purpose models to control noisy translations may not be necessary because models learnt using entity annotations are robust to noise in Q&A data.

Instead of using noisy Q&A data, new approach (Bernhard and Gurevych, 2009) to build parallel corpus from reliable sources has showed improvements. They proposed to use as a parallel training data comprising of set the definitions and glosses provided for the same term by different lexical semantic resources. We move beyond terms and capture relationships between entities and terms using the page contents and link structure in Wikipedia.

Apart from translation models there are other approaches (Gao et al., 2004) that try to extend the existing language models for adhoc retrieval by incorporating term relationships or dependencies. Some expand queries using word relationships derived from co-occurrence thesaurus (Bai et al., 2005; Qiu and Frei, 1993), hand-crafted thesaurus (Liu et al., 2004; Voorhees, 1994) and combination of both (Cao et al., 2005).

(b) *Incorporation of query context information in translation models using latent factor modeling and smoothing approaches:* All these existing approaches mentioned above are considered to be context independent, in that they do not take into account any contextual information in modeling word word relationships. Topic signature model (Zhou et al., 2007) exploited contextual information

[11] $\alpha_{qa}^{(e,e')} = (1 - 0.95)$

by decomposing a document into a set of weighted topic signatures and use it for model smoothing. This model turns out to be inefficient when confronted with ambiguous words and phrases because it is unable to disambiguate the sense of topic signatures. Others (Liu and Croft, 2004) perform semantic smoothing by means of clustering. Recently (Tu et al., 2010; Cai et al., 2011; Zhou et al., 2011) showed improved performance by performing sense based smoothing for document retrieval, latent topic mining and phase based retrieval respectively. Contrary to these approaches we used entity disambiguation to capture contextual information for improving Q&A retrieval.

(c) *Complementary ideas for improving retrieval performance that can be used alongside translation models:* Other work on question retrieval include the use of category information available (Cao et al., 2010), learning-to-rank techniques (Bian et al., 2008; Surdeanu et al., 2008; Bunescu and Huang, 2010), proposed a syntactic tree matching ((Wang et al., 2009) or question structure for important phrase matching (Duan et al., 2008)). These methods seem orthogonal to ours, in some cases complementary and can be leveraged to get an even better performance

There also exists work where exploiting entity based representation has been found helpful in information retrieval (Singh et al., 2009; Egozi et al., 2011; Meij et al., 2008; Grootjen and van der Weide, 2006). In our work we use entity annotations in Q&A retrieval context. There is also some work on using Wikipedia in general web search (Xu et al., 2009).

6 Conclusion

In this work we extend word based model to incorporate semantic concepts for addressing the lexical gap issue in retrieval models for large online Q&A collections. Compared to the existing translation based model, our model is more robust and effective in that it can perform context aware expansions. We proposed ways to embed rich information freely available in Wikipedia into our models and combine it one learnt from Q&A corpus. Experiments performed on a large real Q&A data demonstrate that all configurations of ETLM significantly outperforms existing models for Q&A retrieval.

Acknowledgments

Thanks to Srujana Merugu for helpful discussions. Thanks to the anonymous reviewers for helping us improve the presentation.

References

Lada A. Adamic, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. 2008. Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 665–674, New York, NY, USA. ACM.

Jing Bai, Dawei Song, Peter Bruza, Jian-Yun Nie, and Guihong Cao. 2005. Query expansion using term relationships in language models for information retrieval. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 688–695, New York, NY, USA. ACM.

Adam Berger and John Lafferty. 1999. 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. ACM.

Delphine Bernhard and Iryna Gurevych. 2009. Combining lexical semantic resources with question & answer archives for translation-based answer finding. In *ACL-IJCNLP '09: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*, pages 728–736, Morristown, NJ, USA. Association for Computational Linguistics.

Jiang Bian, Yandong Liu, Eugene Agichtein, and Hongyuan Zha. 2008. Finding the right facts in the crowd: factoid question answering over social media. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 467–476, New York, NY, USA. ACM.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19:263–311, June.

Razvan Bunescu and Yunfeng Huang. 2010. Learning the relative usefulness of questions in community qa. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP

'10, pages 97–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

R. Bunescu and M. Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of EACL*, volume 6, pages 9–16.

Li Cai, Guangyou Zhou, Kang Liu, and Jun Zhao. 2011. Learning the latent topics for question retrieval in community qa. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 273–281, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Guihong Cao, Jian-Yun Nie, and Jing Bai. 2005. Integrating word relationships into language models. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 298–305, New York, NY, USA. ACM.

Xin Cao, Gao Cong, Bin Cui, and Christian S. Jensen. 2010. A generalized framework of exploring category information for question retrieval in community question answer archives. In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 201–210, New York, NY, USA. ACM.

Huizhong Duan, Yunbo Cao, Chin yew Lin, and Yong Yu. 2008. Searching questions by identifying question topic and question focus. In *In Proceedings of 46th Annual Meeting of the Association for Computational Linguistics: Human Language Tchnologies (ACL:HLT)*.

Ofer Egozi, Shaul Markovitch, and Evgeniy Gabrilovich. 2011. Concept-based information retrieval using explicit semantic analysis. *ACM Trans. Inf. Syst.*, 29(2):8:1–8:34, April.

Paolo Ferragina and Ugo Scaiella. 2010. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1625–1628, New York, NY, USA. ACM.

Jianfeng Gao, Jian-Yun Nie, Guangyuan Wu, and Guihong Cao. 2004. Dependence language model for information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 170–177, New York, NY, USA. ACM.

F. A. Grootjen and Th. P. van der Weide. 2006. Conceptual query expansion. *Data Knowl. Eng.*, 56(2):174–193, February.

Jiwoon Jeon, W. Bruce Croft, and Joon Ho Lee. 2005. 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. ACM.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 133–142, New York, NY, USA. ACM.

Thorsten Joachims. 2006. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 217–226, New York, NY, USA. ACM.

S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti. 2009. 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.

Jung-Tae Lee, Sang-Bum Kim, Young-In Song, and Hae-Chang Rim. 2008. 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. Association for Computational Linguistics.

Xiaoyong Liu and W. Bruce Croft. 2004. Cluster-based retrieval using language models. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 186–193, New York, NY, USA. ACM.

Shuang Liu, Fang Liu, Clement Yu, and Weiyi Meng. 2004. An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 266–272, New York, NY, USA. ACM.

Xunying Liu, Mark J. F. Gales, and Philip C. Woodland. 2008. Context dependent language model adaptation. In *INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, September 22-26, 2008*, pages 837–840. ISCA.

Olena Medelyan, David Milne, Catherine Legg, and Ian H. Witten. 2009. Mining meaning from wikipedia. *Int. J. Hum.-Comput. Stud.*, 67:716–754, September.

Edgar Meij, Dolf Trieschnigg, Maarten de Rijke, and Wessel Kraaij. 2008. Parsimonious concept modeling. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 815–816, New York, NY, USA. ACM.

R. Mihalcea and A. Csomai. 2007. Wikify!: linking

documents to encyclopedic knowledge. In *CIKM*, volume 7, pages 233–242.

D. Milne and I.H. Witten. 2008. Learning to link with wikipedia. In *Proceeding of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM.

Robert C. Moore. 2004. Improving ibm word-alignment model 1. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yonggang Qiu and Hans-Peter Frei. 1993. Concept based query expansion. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, pages 160–169, New York, NY, USA. ACM.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1375–1384, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Amit Singh and Karthik Visweswariah. 2011. CQC: classifying questions in cqa websites. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 2033–2036, New York, NY, USA. ACM.

Amit Singh, Sayali Kulkarni, Somnath Banerjee, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Curating and searching the annotated web. In *In SIGKDD Conference, 2009*. ACM.

Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2008. Learning to rank answers on large online qa collections. In *In Proceedings of the 46th Annual Meeting for the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT*, pages 719–727.

Xinhui Tu, Tingting He, Long Chen, Jing Luo, and Maoyuan Zhang. 2010. Wikipedia-based semantic smoothing for the language modeling approach to information retrieval. In *Proceedings of the 32nd European conference on Advances in Information Retrieval*, ECIR'2010, pages 370–381, Berlin, Heidelberg. Springer-Verlag.

Ellen M. Voorhees. 1994. Query expansion using lexical-semantic relations. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 61–69, New York, NY, USA. Springer-Verlag New York, Inc.

Kai Wang, Zhaoyan Ming, and Tat-Seng Chua. 2009. A syntactic tree matching approach to finding similar questions in community-based qa services. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 187–194, New York, NY, USA. ACM.

Yang Xu, Gareth J.F. Jones, and Bin Wang. 2009. Query dependent pseudo-relevance feedback based on wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 59–66, New York, NY, USA. ACM.

Xiaobing Xue, Jiwoon Jeon, and W. Bruce Croft. 2008. 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. ACM.

Chengxiang Zhai and John Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, April.

Xiaohua Zhou, Xiaohua Hu, and Xiaodan Zhang. 2007. Topic signature language models for ad hoc retrieval. *IEEE Trans. on Knowl. and Data Eng.*, 19(9):1276–1287, September.

Guangyou Zhou, Li Cai, Jun Zhao, and Kang Liu. 2011. Phrase-based translation model for question retrieval in community question answer archives. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 653–662, Stroudsburg, PA, USA. Association for Computational Linguistics.

Justin Zobel and Alistair Moffat. 2006. Inverted files for text search engines. *ACM Comput. Surv.*, 38, July.