Harnessing Semantics for Answer Sentence Retrieval

Ruey-Cheng Chen, Damiano Spina, W. Bruce Croft†, Mark Sanderson, Falk Scholer

RMIT University † University of Massachusetts Amherst
{ruey-cheng.chen, damiano.spina}@rmit.edu.au, croft@cs.umass.edu
{mark.sanderson, falk.scholer}@rmit.edu.au

ABSTRACT

Finding answer passages from the Web is a challenging task. One major difficulty is to retrieve sentences that may not have many terms in common with the question. In this paper, we experiment with two semantic approaches for finding non-factoid answers using a learning-to-rank retrieval setting. We show that using semantic representations learned from external resources such as Wikipedia or Google News may substantially improve the quality of top-ranked retrieved answers.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Answer finding; sentence retrieval; non-factoid question answering; semantic representations

1. INTRODUCTION

In this paper, we describe an experiment on using semantics to assist information seeking in a fine-grained retrieval setting. We focus on a relatively new and perhaps more difficult retrieval task called answer passage retrieval [7], a specialized question answering task that looks for non-factoid, multiple-sentence answers from the Web, e.g., *How can UK families become more energy efficient in the future?*. This type of question, also known as non-factoid questions, usually solicits slightly more detailed descriptions or opinions about certain topic subjects. These questions are commonly posted on community question answering sites requesting human input. The answers would be difficult to obtain elsewhere due to either the lack of availability of information, or the limitation of search technology.

Many previous efforts exploited user-generated content to provide reasonable or similar answers to the question being asked [6, 12, 14]. Limiting the scope of search to such well-prepared content can greatly improve answer quality, but this approach benefits only a small set of questions and does not fully exploit the potential of having the entire Web as a backend knowledge base. Considering the amount of human effort involved in developing quality answers, techniques that retrieve answers directly from Webpages could be more scalable and cost efficient [7, 15].

Nevertheless, such an answer-passage retrieval setting is inherently disadvantageous since short text units are more likely to suffer from query mismatch. Furthermore, the so-called "lexical chasm", meaning that the same thought gets radically different wordings at question and answer sides [1], would only make this already severe mismatch problem even harder to deal with. In this case, conventional retrieval models such as TF-IDF or language models could be insufficient because the answer may not necessarily contain any matches with the user's query. We believe that the problem can be alleviated by incorporating a layer of semantic representation into the retrieval process, allowing easier association between sentences about similar topics.

As a first attempt, in this study we consider two approaches of creating sentence-level semantic representations: Explicit Semantic Analysis (ESA) [4], and Word2Vec [11]. ESA is a way to represent text fragments as vectors over human-defined concepts, and it has seen some successful applications in information retrieval and text categorization. Word2Vec is a more up-to-date technology that learns distributed vector representations of words from large amounts of text data. This approach has attracted much attention from research communities because it has an interesting property ("additive compositionality") that allows complex ideas to be expressed by more elementary semantic components. In Sections 3 and 4, both approaches will be explained in more detail.

As part of the evaluation, we build this work around an existing feature-based approach for answer sentence selection. In Section 2 we recount some past work in non-factoid question answering and sentence ranking in general. Details about the two "semantic features" we used, ESA and Word2Vec, are given in Sections 3 and 4. We then describe how to set up the experiment in Section 5. Analysis on the findings is given in Section 6. We conclude this work in Section 7.

2. RELATED WORK

Berger et al. [1] pioneered the task of finding answers by exploring a wide range of probabilistic models such as query expansion, statistical translation, and latent variable mod-

ESAIR'15, October 23, 2015, Melbourne, Australia.

DOI: http://dx.doi.org/10.1145/2810133.2810136.

els. The idea of using translation models to retrieve semantically similar sentences was further exploited in many later efforts on answer retrieval [6, 13, 14, 16]. The research trend by then was to retrieve answers from specific web resources such as question answering archives [6, 16] and FAQ pages [13] rather than from general webpages. Webpages may also contain answers, but retrieving them can take extra effort since the content may adhere less to the query topic, or spread across multiple paragraphs.

Metzler and Kanungo [9] revisited the sentence retrieval problem from a machine learning perspective and developed a learning-to-rank approach on top of TREC Novelty Track data. But still the data is not entirely focused on direct answer finding in webpages. Due to the lack of a dedicated benchmark, Keikha et al. [7] set out to develop their own annotations on top of the GOV2 collection, reusing previous TREC topics. They tested several retrieval and query expansion settings, and showed that true answers tend to have distinctively lower scores than top-ranked passages. Their result suggests that the conventional notion of topical relevance is ineffective on this type of task.

Some previous work has also looked into the problem of answer ranking in community question answering services. Surdeanu et al. [14] put together a sizable test collection from Yahoo! Answers, and tackled the answer ranking problem with a learning-to-rank approach that integrates many similarity and translation features. Some recent efforts along this line focus on features such as discourse relation [5] and distributed word representations [12]. Note that even though the problem of answer ranking is closely related to answer finding, the size of text units (i.e., texts can span multiple paragraphs) and the depth of the retrieved set (i.e., questions with more than 20 answers are scarce) are entirely different. It would be incorrect to assume that these two tasks are the same.

3. EXPLICIT SEMANTIC ANALYSIS

Gabrilovich and Markovitch proposed Explicit Semantic Analysis (ESA) [4], a method that leverages vast amount of common-sense knowledge on the Internet to compute semantic relatedness of arbitrary texts. The key idea is to represent texts as "a weighted mixture of a predetermined set of *natural* concepts." Here, the natural concepts are usually referred to as page entries on Wikipedia. Semantic relatedness is then computed on top of this representation, with usual vector similarity measures such as the cosine similarity.

As short text segments can usually benefit from this enriched representation, we expect that this transformation could bring questions closer to the respective answers. For instance, running the description of Topic 830 "*Locate past or present model railroad layouts*" through ESA and taking only the top 10 concepts would yield the following representation, which is precise and rich in semantics:

Model Railroader	0.0074
John Armstrong (model railroader)	0.0065
John Whitby Allen	0.0061
San Diego Model Railroad Museum	0.0058
Great Northern Depot (Wayzata, Minnesota)	0.0057
Frank Ellison	0.0056
SE&CV	0.0056
Rail transport modelling	0.0055
Virginian and Ohio	0.0054
Gorre & Daphetid	0.0052

The ESA representation for an arbitrary text can actually be created by simply running the text as a query to an index over all Wikipedia pages. The retrieved top-k page IDs are deemed as associated concepts and the respective TF-IDF scores as weights in the vector representation. Despite TF-IDF being exclusively used in the original paper, we found that its role can practically be taken by any reasonable similarity function. In our implementation, we use language model retrieval approach with Dirichlet smoothing to compute the weights.

4. WORD2VEC

Mikolov et al. proposed an efficient method for learning vector representations of words from large amounts of unstructured text data, called the Skip-gram model [10]. Some extensions and data for building the Skip-gram model are released as an open-source project under the name of Word2Vec.[1] In this learned vector space, words with similar meanings are represented as vectors in close distance. Hence, they can be compared to measure semantic similarity—for instance, by using cosine similarity. It is also interesting to note that the learned vectors may in some degree exhibit additive compositionality, i.e., meanings can be added together to express complex ideas. With some concrete examples, Mikolov et al. concluded that the learned representations "exhibit a linear structure that makes it possible to perform precise analogical reasoning using simple word vector arithmetic" [11]. Although it relies on a neural network architecture to compute word vectors, training the Skip-gram model appears to be efficient.

5. EXPERIMENTS

We first describe our experimental setup in Section 5.1, and study the performance of individual features on the answer finding task in Section 5.2. Then, in Section 5.3 we look into the problem of integrating all these features into a learning-to-rank framework.

5.1 Setup

Data. We conducted experiments using a non-factoid answer retrieval dataset derived from the GOV2 test collection provided by Keikha et al [7]. To prepare the data, a set of query topics from Topics 701–850 were first identified as having passage-level answers, and then annotations were created for each of these topics on the top 50 retrieved documents. The produced dataset contains 8,027 manually annotated answer passages to 82 GOV2 query topics with graded judgments on 4 relevance levels.

Learning-to-Rank. We considered different learning-to-rank algorithms in this experiment: Linear Regression, Multiple Additive Regression Trees (MART) [3] and Coordinate Ascent [8]. We used the RankLib[2] implementations for all these algorithms. In our setting, we use 5-fold cross validation, including a 20% validation split inside the training set, and using NDCG@10 as the objective function. For each feature/algorithm combination, this cross validation procedure is repeated 10 times over randomized partitions to reduce

[1] https://code.google.com/p/word2vec
[2] http://sourceforge.net/p/lemur/wiki/RankLib/

Table 1: Performance of single features, with statistical significance tested against features LanguageModel (*p < 0.05, **p < 0.01) and ESACosineSimilarity (†p < 0.05, ‡p < 0.01). Best results are in boldface.

Feature Set	Feature	NDCG@10	P@10	MRR
Metzler-Kanungo (MK)	SentenceLength	0.0036**‡	0.0037**‡	0.0174**‡
	SentenceLocation	0.0000**‡	0.0000**‡	0.0056**‡
	ExactMatch	0.0194**‡	0.0220**‡	0.0529**‡
	TermOverlap	0.0618	0.0622†	0.1978
	SynonymOverlap	0.0272*‡	0.0293**‡	0.1058‡
	LanguageModel	0.0721	0.0866	0.1980
Semantic Features	ESACosineSimilarity	**0.1053**	**0.1171**	**0.2690**
	Word2Vec	0.0634†	0.0720	0.1924

the effect of data split. We report the average performance in this paper.

Metzler-Kanungo Features. We include Metzler and Kanungo's features [9] as they are straightforward to implement and their work was motivated for a similar task. In their experiment, Metzler and Kanungo used the following 6 features:

SentenceLength Number of terms in a sentence, after stopwords are removed.

SentenceLocation Position of a sentence, normalized by the number of sentences in that document.

ExactMatch Equals 1 if there is an exact lexical match of the query string occurs in the sentence, and 0 otherwise.

TermOverlap Fraction of query terms that occur in the sentence, with stopping and stemming.

SynonymOverlap Fraction of query terms that either occur or have a synonym in the sentence, with stopping and stemming. In our experiment we used WordNet synsets in NLTK[3] to generate synonyms.

LanguageModel Log likelihood of the query terms being generated from the sentence language model with Dirichlet smoothing [17]:

$$\sum_{t \in Q} tf_{t,Q} \log \frac{tf_{t,S} + \mu P(t|C)}{|S| + \mu} \tag{1}$$

We created a Galago[4] index over the entire GOV2 collection to serve as the background model $p(\cdot|C)$.

Hereafter in our evaluation, this set of features is referred to as the Metzler-Kanungo (MK) set. We made a few changes in our implementation: (i) we used Porter stemming and InQuery stoplist throughout, (ii) we optimize the performance of the language model beforehand through a grid search over μ, (iii) the value of the language model is normalized (by subtraction) against its mean within individual queries. The last change has no impact on retrieval performance but allows machine learning algorithms to pick up the within-query effect more easily.

[3] http://www.nltk.org

[4] http://www.lemurproject.org/galago.php

Semantic Features. Adding to the top of the MK feature set are our proposed *semantic* features.

ESACosineSimilarity Cosine similarity between the query ESA vector and the sentence ESA vector. We used a recent dump of English Wikipedia (June 2015) to generate ESA representations for the query and all the sentences. For efficiency, we consider only the top 100 concepts when constructing such ESA vectors.

Word2Vec Average pairwise cosine similarity between any query-word vector $\vec{u}$ and any sentence-word vector $\vec{v}$:

$$\frac{1}{|Q||S|} \sum_{\vec{u} \in Q} \sum_{\vec{v} \in S} \frac{\vec{u} \cdot \vec{v}}{\|\vec{u}\|\|\vec{v}\|} \tag{2}$$

Note that when all word vectors are unit vectors, the average pairwise cosine similarity is equivalent to the dot product of the mean query vector and the mean sentence vector. We used *gensim*[5] to generate this feature. The pre-trained distributional model we used [10], which is based on the text contents from Google News, amounts to roughly 100 billions words.

Evaluation Metrics. In answer finding, we mainly focus on top-ranked retrieval performance. Thus, we report Normalized Discounted Cumulative Gain at top 10 (NDCG@10) using logarithmic discount, Precision at 10 (P@10) and Mean Reciprocal Rank (MRR). We use a two-tailed pairwise t-test with Bonferroni correction to test for statistical significance. Significant improvements are denoted as **,‡ for $\alpha = 0.01$ and *,† for $\alpha = 0.05$, respectively.

5.2 Single Feature Ranking

Before tackling the learning-to-rank task, we first test the effectiveness of individual features, i.e., rank the sentences according to the given feature values. Table 1 shows the results obtained for both the MK and the semantic feature sets. The result shows that LanguageModel is the best feature in the MK feature set, with TermOverlap being the second. It is statistically significantly better than most of the other features. The result is generally in line with Metzler and Kanungo [9] in the task of sentence selection for query-biased summarization.

For the semantic features, average pairwise cosine similarity over Word2Vec vectors achieves comparable performance

[5] http://radimrehurek.com/gensim

Table 2: Performance of Learning-to-Rank approaches using different feature sets and machine learning algorithms. Best results are printed in boldface. For each algorithm, statistical significance is tested against the MK ($^{*}p < 0.05$, $^{**}p < 0.01$) and MK + ESACosineSimilarity ($^{\dagger}p < 0.05$,$^{\ddagger}p < 0.01$) feature sets.

Feature Set	Algorithm	NDCG@10	P@10	MRR
MK	Linear Regression	0.0792	0.1016	0.1998
MK + ESACosineSimilarity		0.0754	0.0935	0.2019
MK + Word2Vec		0.0724	0.0900	0.1882
All Features		**0.1123**	**0.1254**	0.2637
MK	Coordinate Ascent	$0.0667^{\dagger}$	$0.0788^{\dagger}$	0.1954
MK + ESACosineSimilarity		0.1080^{*}	0.1221^{*}	0.2694
MK + Word2Vec		0.0810	0.0936	0.2278
All Features		0.1114^{**}	0.1240^{*}	**0.2778**
MK	MART	$0.0603^{\ddagger}$	$0.0699^{\dagger}$	0.1754
MK + ESACosineSimilarity		0.0994^{**}	0.1119^{*}	0.2404
MK + Word2Vec		0.0699	0.0769	0.1985
All Features		0.0953^{*}	0.1088^{**}	0.2363

to the best two features in the MK set. The other feature, ESA, obtains the best results for all the reported evaluation measures, although the difference with respect to Language-Model is not significant. The result on MRR suggests that, on average, ESA is capable of delivering a relevant answer within the first 4 retrieved sentences. The proposed semantic features perform similarly or better than the features considered in previous work, suggesting that the external resources such as Wikipedia or Google News (where the pre-trained distributional models are derived) can be used to easily improve non-factoid question answering systems.

5.3 Learning-to-Rank

In this section, we investigate whether the proposed features provide auxiliary signals of relevance that complement the function of the MK feature set. Table 2 shows the performance of combining different feature sets for Linear Regression, Coordinate Ascent and MART. The result shows that combining either ESACosineSimilarity or Word2Vec with the MK feature and trained with Coordinate Ascent or MART performs better than not using any of the two—although the absolute improvement is not big. For instance, the ranker trained by Coordinate Ascent on the MK and ESACosineSimilarity features obtains an NDCG@10 score of 0.1080^{*}, which corresponds to a 62% of relative improvement with respect to the baseline performance.

Except for the case of MART, combining all features gives the best performance. In particular, the Linear Regression ranker on all features obtains the best—but not significant—overall NDCG@10 and P@10 results. Besides, Coordinate Ascent obtains similar results, being statistically different ($\alpha = 0.05$) with respect to using the MK features alone. The result in MRR shows that, on average, this configuration is capable of retrieving a correct answer sentence within the first 4 attempts.

In summary, the use of external resources such as knowledge bases and distributional models can adds complementary signals to an established set of features based on largely lexical matching and synonyms. The semantic features specifically tailored for our experiment could also be useful in other type of ranking task.

5.4 Per-Topic Performance

We look further into the per-topic performance between experimental runs. Figure 1 shows the per-topic differences between individual runs and the MK baseline. We found that less than 50% of query topics benefit from the MK-ESA combination. The same goes for all features, and for the MK-Word2Vec combination only a small set of topics saw improvements. Despite the small improvements, retrieval performance seems to add up after combining both ESA and Word2Vec features. It is interesting to note that, in the rightmost panel, many topics with zero improvement on all features actually failed on the MK baseline in bringing any answer to top 10. In about a third of cases where the baseline failed (zero in NDCG@10 as denoted by diamond-shaped markers), the all features run is capable of retrieving relevant answers in the top set. Overall, using semantic features in answer sentence retrieval can have a mixed result, but in general positive effects are more common than negative ones.

6. ANALYSIS

Our experimental result on combining features suggests that ESA cosine similarity and Word2Vec tend to retrieve different sets of answers and therefore may complement each other in the task of answer finding. To investigate how, we look into top-ranked sentences produced for individual queries and make comparisons between the experimental runs. For ease of making contrasts, we deliberately leave out the "All Features" run. For each run, we pull ranking results from one of the shuffled cross-validation runs trained by Coordinate Ascent. We focus on query topics for which one or two experimental runs succeeded, whereas others did not, in bringing an answer sentence to the top.

Three such examples topics are given in Table 3. As one might expect, the MK baseline is in favor of sentences that match better with the query, and matching on less discriminative terms, such as common terms, is considered less important—much like what we get from a fine-tuned retrieval model. This could be a disadvantage when the focus or head concept of a given question is not precise, or even ambiguous, e.g., "*security measures*" in Topic 711 and "*state of . . . relations*" 770.

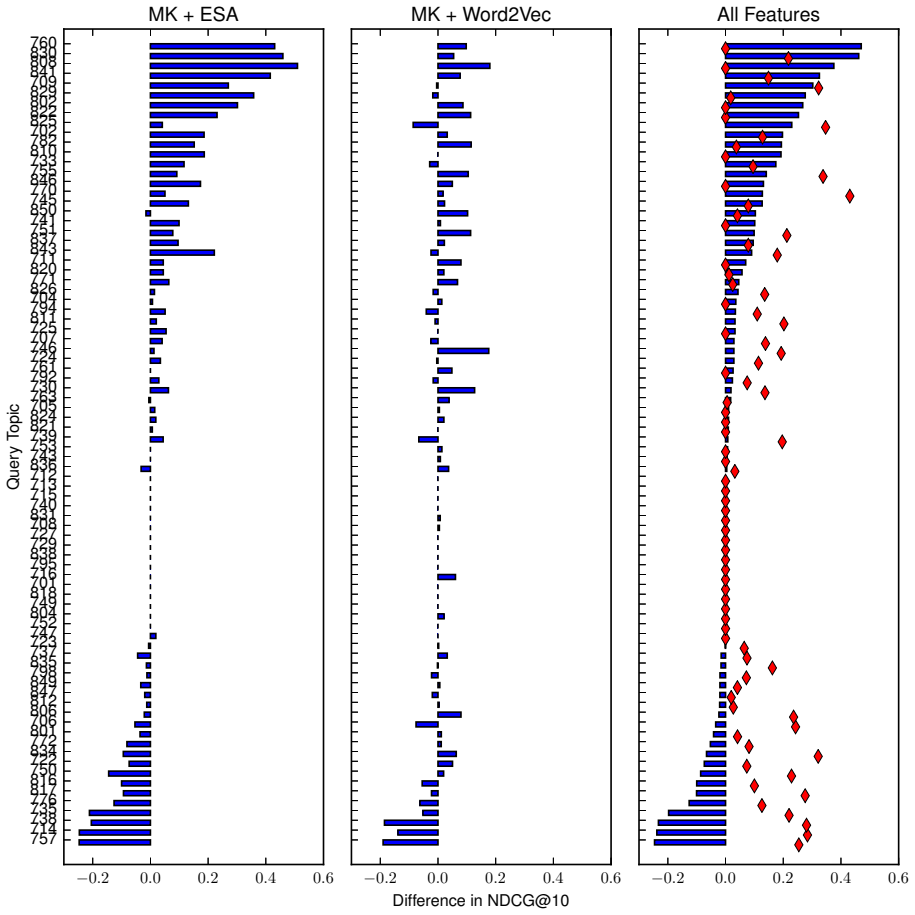

Figure 1: Per-topic difference in NDCG@10 against baseline for all experimental runs, with topics sorted according to the rightmost panel ("All Features"). Diamond-shaped markers on the right indicate the baseline performance (x-axis in this case indicates absolute NDCG@10 scores).

By introducing semantic features, we bias the retrieval process towards slightly different directions and hopefully this would mitigate the issue with the MK baseline. With ESA cosine similarity, we found that many retrieved answers are informative and topically coherent. They tend to cover some specific aspects in the question instead of the whole query context, a side effect of doing query-sentence matching in a space of concept surrogates. One obvious problem is that certain semantic structures cannot be appropriately represented by lexical matching to knowledge base concepts, e.g., "*... employed at train station*" in Topic 711 and "*How are ... acquired?*" in Topic 817. So it is fair to say that retrieval using ESA cosine similarity still largely depends on lexical matching, though done in a way that takes more contexts into account.

The result with the Word2Vec feature appears more structurally coherent to the original question, because pairwise similarity tends to retrieve sentences with equal emphasis on each query term. Although Word2Vec produces sentences where term-wise topics seem more balanced, this does not always lead to coherent answers. Again, the influence of head concept in the original query gets washed away by taking pairwise similarity over query and sentence words, so there is no guarantee that the main information need would be fulfilled in retrieved answers (see Topic 770 for example).

We also looked at top-ranked sentences from the respective single-feature runs of ESA and Word2Vec, on query topics where both ESA and Word2Vec succeeded in bringing answers to top 10. In Table 4, we give two such topics with results from both runs with some more details. In Topic 782, ESA retrieves some shipping lists of oranges but failed to boost results about seasonal fruits. For the same topic, Word2Vec does well in promoting seasonal, orange-related results—given the fact that those concepts are close to each other in the distributional semantic space. However, it seems too keen on lexically matching the query term "*varieties*." Certainly, none of the semantic features recognize *varieties* as a cue of specific (e.g., list) question type, but ESA appears to suffer less due to the nature of concept expansion in a knowledge base. In Topic 841, ESA leverages the knowledge encoded in an established Wikipedia concept "Camel" to generate a list of relevant prehistoric animals in North America. On the other hand, Word2Vec does not stick to the head concept all the time and sometimes produces off-topic results. Also, Word2Vec appears vulnerable

Table 3: Some example query topics for which individual rankers select different top-1 sentences. Rel indicates relevance level. Lexically matched sentence terms are underlined.

Run	Rel	Top-1 Sentence
711: *What security measures have been employed at train stations due to heightened security concerns?*		
MK	0	The biggest concern in the minds of security personnel is the possibility of a person boarding a bus or train with a gun or other weapon.
MK+ESA	0	Two major cooperatives in the fertilizer industry, Farmland and CF Industries,have always been aware of potential security concerns, but both have increased their guard as security threats have become a heightened concern in a post-Sept. 11 world.
MK+Word2Vec	3	"Amtrak responded admirably to the crisis, quickly training personnel on heightened security and safety procedures, assigning more security officers to stations and trains, and requiring passengers to bring photo identifications for security checks," Schumer wrote.
770: *What is the state of Kyrgyzstan-United States relations?*		
MK	0	(3) Kyrgyzstan concluded a bilateral investment treaty with the United States in 1994.
MK+ESA	4	The extension of unconditional normal trade relations treatment to the products of Kyrgyzstan will enable the United States to avail itself of all rights under the World Trade Organization with respect to Kyrgyzstan.
MK+Word2Vec	0	(begin text) U.S. DEPARTMENT OF STATE Office of the Spokesman January 15, 2002 Media Note RELIGIOUS LEADERS FROM KYRGYZSTAN EXAMINE ISLAM IN THE UNITED STATES
817: *How are naming rights to sports stadiums acquired?*		
MK	3	Recently in the United States, naming rights for new professional sports stadiums typically have yielded $2 million to $2.5 million per year for terms of ten to thirty years.
MK+ESA	0	The analogous data from the estimated value of naming rights for sports stadiums indicate that naming rights represent a small portion of total facility costs.
MK+Word2Vec	3	Naming rights have migrated from sports stadiums and arenas to performing arts centers.

to spamming (see the 3rd sentence at Row 4 in Table 4) due to the way sentence-level similarity is computed.

7. CONCLUSION

Answer sentence retrieval is a sophisticated task since the answers may not necessarily contain any lexical matches with the query and cannot be easily captured using extraction techniques in question answering. Nevertheless, in this paper we demonstrate that the difficulty in finding answers can be leveraged in a feature-based framework by incorporating *semantic* features that make use of external resources to increase the chance of finding correct answers that may have small lexical similarity with the given question.

In particular, our preliminary experiments show a promising direction to explore which consists of using knowledge bases such as Wikipedia or pre-trained distributional semantic models such as Word2Vec.

As immediate future work, we plan to define new features by using entity linking systems [2] and explicitly representing semantics at phrase or sentence level by harnessing the compositionality property of Word2Vec.

8. ACKNOWLEDGMENTS

This research is supported in part by ARC Discovery Grant DP140102655, in part by ARC Project LP130100563, in part by Real Thing Entertainment Pty Ltd, and in part by NSF Grant IIS-1419693.

9. REFERENCES

[1] A. Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal. Bridging the Lexical Chasm: Statistical Approaches to Answer-finding. In *Proceedings of SIGIR'00*, pages 192–199, 2000.

[2] P. Ferragina and U. Scaiella. Fast and accurate annotation of short texts with wikipedia pages. *Software, IEEE*, 29(1):70–75, 2012.

[3] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[4] E. Gabrilovich and S. Markovitch. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of IJCAI'07*, pages 1606–1611, 2007.

[5] P. Jansen, M. Surdeanu, and P. Clark. Discourse Complements Lexical Semantics for Non-factoid Answer Reranking. In *Proceedings of ACL'14*, pages 977–986.

[6] J. Jeon, W. B. Croft, and J. H. Lee. Finding Similar Questions in Large Question and Answer Archives. In *Proceedings of CIKM'05*, pages 84–90, 2005.

[7] M. Keikha, J. H. Park, W. B. Croft, and M. Sanderson. Retrieving Passages and Finding Answers. In *Proceedings of ADCS'14*, pages 81:81–81:84, 2014.

[8] D. Metzler and W. B. Croft. Linear feature-based models for information retrieval. *Information Retrieval*, 10(3):257–274, 2007.

Table 4: Top-ranked sentences for Topics 782 and 841, produced respectively by single-feature ESA (top) and single-feature Word2Vec (bottom).

782: *What are the varieties of oranges and when is each in season?*	
3	Expected gift fruit shipments under the 6-R program, and non-certified usage, 1999-00 season Type 1,000 boxes Early and Midseason Oranges 1,500 Valencia Oranges 700 White Seedless Grapefruit 400 Colored Seedless Grapefruit 800 Temples 100 Tangelos 200 Tangerines 200 K-Early Citrus Fruit 5
0	California's lemons, like its oranges, are smaller so far this season.
0	Bearing acreage declined by 2,000 for Navel oranges but was the same as the previous season for other citrus varieties.
2	Expected gift fruit shipments under the 6-R program, and non-certified usage, 2000-01 season Type 1,000 boxes Early and Midseason Oranges 2,000 Valencia Oranges 1,000 White Grapefruit 500 Colored Grapefruit 1,000 Temples 100 Tangelos 200 Tangerines 300 K-Early Citrus Fruit 5
0	Hamlins are the second most popular variety, making up 26 percent of the citrus trees produced, and are the leading early season orange.Midsweet is the most utilized midseason orange selection and the third most popular variety.Pineapple midseason orange is the fifth most popular variety.
0	A variety of the Washington Navel orange is the principal orange product of Texas.
4	The Moro orange (a type of blood orange) and the red Cara Navel are two western-grown seasonal varieties.
0	Because of significant differences in fruit set, size, drop, and harvest patterns of this variety from other oranges, the Navel orange forecast is computed separately from the other oranges and is used as an add-on indicator in the early-mid and all orange forecasts.
841: *Provide information on camels in North America in both prehistoric and modern times.*	
4	These camels became extinct in North America several thousand years ago.
0	Pleistocene mega fauna included "grazing herds of elephants, mammoths, rhinos, camels, horses, burros, ground sloths" and others including prehistoric cattle.
3	These early cultures pursued dwindling herds of Pleistocene mammoth, camel, horse, bison and other now-extinct species into most of North and South America.
0	There is some debate about what caused the extinction of over 30 species, including the mastodon, mammoth, horse, camel, giant sloth, short faced bear, dire wolf, giant beaver, and long-horned bison, most of which were larger than their modern day cousins.
3	Ancient camels and horses originated on our continent and migrated to Asia, leaving only fossils in North America.
0	Prehistoric Horses in North America The fossil record indicates that horses first evolved in North America about 60 million years ago and from there spread to other continents (Denhardt 1975).
0	Smith - public archeology, archeological resource protection, prehistoric archeology, zooarcheology, museology, Arctic, Subarctic, southeastern North America Audrey M. Trauner - public archeology, prehistoric and historical archeology, museology, zooarcheology, southeastern North America Robert C. Wilson - CRM, historic and prehistoric archeology, remote sensing, geographic information systems, database management, southeastern North America
0	SUMMARY: The cultural resources of the Upper Santa Cruz subarea are the product of thousands of years of human settlement from the earliest prehistoric times to the modern day.

[9] D. Metzler and T. Kanungo. Machine Learned Sentence Selection Strategies for Query-Biased Summarization. In *SIGIR Learning to Rank Workshop*, 2008.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*, Jan. 2013. arXiv: 1301.3781.

[11] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[12] P. Molino and L. M. Aiello. Distributed Representations for Semantic Matching in non-factoid Question Answering. In *Proceedings of SIGIR Workshop on Semantic Matching in Information Retrieval (SMIR'14)*, pages 38–45, 2014.

[13] S. Riezler, A. Vasserman, I. Tsochantaridis, V. Mittal, and Y. Liu. Statistical Machine Translation for Query Expansion in Answer Retrieval. In *Proceedings of ACL'07*, pages 464–471, 2007.

[14] M. Surdeanu, M. Ciaramita, and H. Zaragoza. Learning to Rank Answers on Large Online QA Collections. In *Proceedings of ACL'08*, pages 719–727, 2008.

[15] S. Verberne, H. v. Halteren, D. Theijssen, S. Raaijmakers, and L. Boves. Learning to rank for why-question answering. *Information Retrieval*, 14(2):107–132, June 2010.

[16] X. Xue, J. Jeon, and W. B. Croft. Retrieval Models for Question and Answer Archives. In *Proceedings of SIGIR'08*, pages 475–482, 2008.

[17] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, Apr. 2004.