3.2 Retrieval Method

Vocabulary mismatch is a problem that particularly affects short-text retrieval and semantic features are an effective way to alleviate this problem. Our retrieval task also demands the use of semantic similarity because we do not seek to find entities that appear exactly as described in the query sentence. We combine the effectiveness of both syntactic and semantic matching for computing sentence similarity. We use BM25 for keyword matching and Sentence Embedding (SE) [15] to compute semantic similarity. We assume that word matching is more important for the user selected words and semantic similarity is important for matching similar entities, and combine the benefits of both to create an effective model.

We capture semantic level matching between a sentence pair to retrieve and score the top k sentences against the query sentence. We use the average of word embedding to obtain sentence embedding for the query and candidate sentences. Word embedding methods learn a low-dimensional vector representation of words from a large, unstructured text corpus; we use the skip-gram model proposed by Mikolov et al. [9] to generate representations for words. Finally, we use cosine similarity to compute similarity between a query and a candidate sentence. Our approach is inspired by Wieting et al. [15], who showed that a simple averaging over the embedding of the words in a sentence provides an effective representation for that sentence and that representation is particularly helpful for sentence similarity task.

For all our ranking techniques, we use the Stanford Named Entity Recognizer [6] to reject candidate sentences that do not contain entities of the appropriate type. While this may introduce false-negatives, it greatly increases precision of our system, and allows our other techniques to focus on ranking and increasing recall.

3.3 Query Expansion (QE)

In order to obtain a broader and generalized representation of query sentence, we use Pseudo Relevance Feedback (PRF) for query expansion at the sentence level. We use BM25 to retrieve PRF sentences given the query sentence and compute the average over the embedding of those sentences to obtain a more robust representation of the query.

3.4 PRF with User Feedback

We make use of the context words selected by user for finding PRF sentences with BM25 technique. We expect that a keyword-based search technique would find sentences focusing on user selected terms. Suppose, our original query sentence Q_o contains a list of n terms, and user u has constructed a list, CW from Q_o, of k words, where each word appears there one or more times. Now, in order to get an expanded query Q_e, we simply concatenate all the terms in CW with Q_o. The goal of this process is to assign term importance in a query by repeating the term multiple times. Even though it is not a sophisticated method of incorporating user-provided term weights, it works well in practice. We search the sentence corpus with Q_e and use the top k retrieved sentences, S_{topk} for obtaining a better representation for Q_o. Finally, we compute the average of the sentence embeddings from the sentences in set $Q_f = \{Q_o \cup S_{topk}\}$, perform SE based search using Q_f and re-rank them using the method described in 3.2.

4 EXPERIMENTAL RESULTS

4.1 Evaluation Metrics and Relevance

We use novelty versions of recall and precision, standard measures modified so that only the first instance of a target entity is considered relevant. We use recall@k to measure the number of relevant (and unique) entities observed in the top k sentences and we use precision@k to measure the proportion of sentences in the top k that contain relevant (and unique) entities. We stress that the relevance of a sentence is determined by two properties: containing a relevant entity *and* being unique in the ranked list so far. We also report MAP@1000 and recall@1000, measures that are important because when we want to perform two-stage retrieval and ranking, retrieving most of the entities in the top 1000 sentences becomes crucial.

4.2 Baseline Methods

We compare the effectiveness of user feedback against three non-interactive baselines.

- **SE** is the Sentence Embedding (SE) based search described in Section 3.2 that assumes no information regarding term importance.
- **SE + PRF** is similar to the approach described in Section 3.4 that uses BM25 to retrieve the top k sentences using the sentence query and then combines those to obtain an expanded query that is used with SE.
- **SE + PRF + CW + Sim (Entity, Token)** is similar to the process of integrating user-selected context words into the query as mentioned in Section 3.4. However, the process of obtaining context word is not based on any human input. We use this baseline to check how better human input is compared to an automatic process that can generate context words. This method computes the similarity of each word in the query sentence with the query entity. Then it uses the five most similar words for performing PRF. Similarity between a word and query entity is computed using the similarity of their embedding. An entity embedding is constructed by the average of the embedding of the words in it.

4.3 Result Discussion

Table 1 summarizes the average performance of term relevance feedback for CNER across the queries that have been annotated by lab and Mechanical Turk participants. Across all forms of feedback, the lab participants created more effective queries than the crowdsource workers: this is reasonable as the lab participants are likely to be more expert searchers.

Overall, term feedback was helpful (10.7% improvement in mAP), and weighted term feedback was even more helpful (14.9% improvement in mAP). The means that our two research questions are both answered positively: user feedback provides improved results for CNER, and allowing users to specify an ordering or weighting on terms is helpful.

In addition, we analyze the impact of adding fewer keywords (and therefore minimizing user involvement). For each of the queries, we selected the top-$k = 1 \ldots 5$ terms based on the weights provided by the users, added them to the original query (with weights). Performance is presented in Figure 3 in terms of Precision@5, Recall@5 and mAP@1000. Although there is some noise, particularly in the

Table 1: Average performance of various methods. Measures are listed in the first row, with high-precision measures listed first. mAP and R are cut-off at depth 1000. The first section of the table presents baselines, then weighted feedback and finally unweighted feedback. The percentage improvement is shown over the Sentence Embedding (SE) baseline.

R@5	R@10	P@5	P@10	mAP	R	Method	Source of Context Words	Weighted?
0.145	0.234	0.180	0.180	0.188	0.891	SE	None	No
0.123	0.183	0.160	0.130	0.153	0.884	SE + PRF	None	No
0.147	0.177	0.200	0.150	0.162	0.865	SE + PRF + CW	Sim (Entity, Word)	No
0.183 (+26.3%)	0.244 (+4.3%)	0.231 (+28.4%)	0.184 (+2.3%)	0.216 (+14.9%)	0.910 (+2.2%)	SE + PRF + CW	Mturk + Lab Participants	Yes
0.204 (+40.7%)	0.267 (+14.2%)	0.237 (31.7%)	0.196 (+8.9%)	0.232 (+23.5%)	0.921 (+3.4%)	SE + PRF + CW	Lab Participants	Yes
0.171 (+18%)	0.229 (-2.2%)	0.222 (+23.4%)	0.176 (-2.3%)	0.205 (+9.1%)	0.900 (+1.1%)	SE + PRF + CW	Mturk Participants	Yes
0.175 (+20.7%)	0.234 (+0.0 %)	0.226 (+25.6%)	0.179 (-0.6%)	0.208 (+10.7%)	0.907 (+1.1%)	SE + PRF + CW	Mturk + Lab Participants	No
0.186 (+28.3%)	0.255 (+9.0%)	0.234 (+30%)	0.194 (+7.8%)	0.220 (+17.1%)	0.920 (+3.3%)	SE + PRF + CW	Lab Participants	No
0.166 (+14.5%)	0.219 (-6.5%)	0.217 (+20.6%)	0.168 (-6.7%)	0.199 (+5.9%)	0.897 (+0.7 %)	SE + PRF + CW	Mturk Participants	No

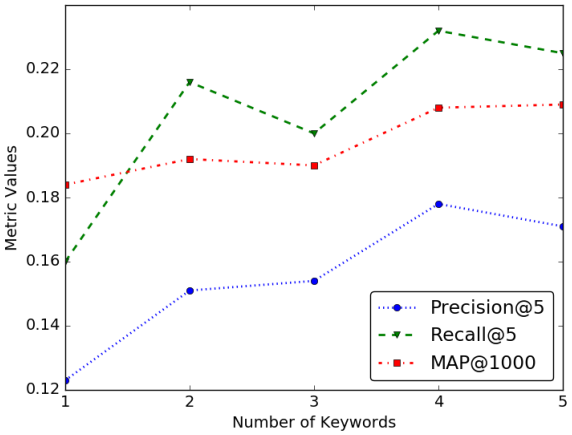

Figure 3: Performance Sensitivity with Keywords Addition

recall of the solution, it is clear that a handful of keywords can be effective (although it does depend on the user and the quality of the terms selected), but more terms do appear to be better.

5 CONCLUSION

We adopt a term relevance feedback technique for list query construction from a sentence and show its effectiveness in entity retrieval. We started this work with two research questions and answered them both affirmatively. We showed that (RQ1) users *can* select better query terms than automatic methods, and that (RQ2) it *is* helpful for the user to identify which terms are best. Our interface for collecting this information was rudimentary and we did not explore alternatives for this study. Future work will look at how an interface can best support a user in providing that information.

ACKNOWLEDGEMENT

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant #IIS-1617408. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

REFERENCES

[1] Krisztian Balog, Pavel Serdyukov, and Arjen P de Vries. 2010. *Overview of the TREC 2010 entity track.* Technical Report. DTIC Document.
[2] N. J. Belkin, C. Cool, D. Kelly, S.-J. Lin, S. Y. Park, J. Perez-Carballo, and C. Sikora. 2001. Iterative Exploration, Design and Evaluation of Support for Query Reformulation in Interactive Information Retrieval. *Inf. Process. Manage.* (2001).
[3] Bhavana Bharat Dalvi, Jamie Callan, and William W. Cohen. Entity List Completion Using Set Expansion Techniques. In *TREC'10.*
[4] Hoa Trang Dang, Jimmy Lin, and Diane Kelly. Overview of the TREC 2006 Question Answering Track. In *TREC'06.*
[5] Gianluca Demartini, Tereza Iofciu, and Arjen P. De Vries. Overview of the INEX 2009 Entity Ranking Track. In *INEX'09.*
[6] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *ACL'05.*
[7] Diane Kelly and Xin Fu. Elicitation of Term Relevance Feedback: An Investigation of Term Source and Context. In *SIGIR'06.*
[8] Jürgen Koenemann and Nicholas J. Belkin. A Case for Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness. In *CHI'96.*
[9] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and Their Compositionality. In *NIPS'13.*
[10] Yael Nemeth, Bracha Shapira, and Meirav Taeib-Maimon. Evaluation of the Real and Perceived Value of Automatic and Interactive Query Expansion. In *SIGIR'04.*
[11] Ian Ruthven. Re-examining the Potential Effectiveness of Interactive Query Expansion. In *SIGIR'03.*
[12] Yaming Sun, Lei Lin, Duyu Tang, Nan Yang, Zhenzhou Ji, and Xiaolong Wang. Modeling Mention, Context and Entity with Neural Networks for Entity Disambiguation. In *IJCAI'15.*
[13] Ellen M Voorhees and Hoa Trang Dang. Overview of the TREC 2005 Question Answering Track. In *TREC'05.*
[14] P.C. Wankat. 2002. *The Effective, Efficient Professor: Teaching, Scholarship, and Service.* Allyn and Bacon.
[15] John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. Towards Universal Paraphrastic Sentence Embeddings. In *ICLR'16.*