The Impact of Semantic Document Expansion on Cluster-based Fusion for Microblog Search

Shangsong Liang, Zhaochun Ren, and Maarten de Rijke

ISLA, University of Amsterdam
{s.liang, z.ren, derijke}@uva.nl

Abstract. Searching microblog posts, with their limited length and creative language usage, is challenging. We frame the microblog search problem as a data fusion problem. We examine the effectiveness of a recent cluster-based fusion method on the task of retrieving microblog posts. We find that in the optimal setting the contribution of the clustering information is very limited, which we hypothesize to be due to the limited length of microblog posts. To increase the contribution of the clustering information in cluster-based fusion, we integrate semantic document expansion as a preprocessing step. We enrich the content of microblog posts appearing in the lists to be fused by Wikipedia articles, based on which clusters are created. We verify the effectiveness of our combined document expansion plus fusion method by making comparisons with microblog search algorithms and other fusion methods.

1 Introduction

Searching microblogs continues to be a challenge, for multiple reasons. For one, the vocabulary mismatch problem takes on a new form. If microblog posts contain only a few words, some of which are misspelled or creatively spelled, the risk of query terms failing to match words observed in relevant posts is large. Additionally, in very short posts, most terms occur only once, making simple operations such as language model estimation difficult [1, 4].

We address these challenges by using data fusion, thereby combining the output of multiple ranking functions, that each combats the unique challenges of microblog search in their own way [2]. That is, instead of searching microblog posts directly, we merge the lists generated by a number of state-of-the-art microblog search algorithms and try to outperform the best component list.

A large number of effective data fusion strategies have been proposed, with the CombSUM family of fusion methods being the oldest and one of the most successful ones in many IR tasks [4, 8]. We are interested in cluster-based fusion [3], as it is a state-of-the-art fusion method that can significantly improve performance in many IR applications [3, 4]. In cluster-based fusion, documents from lists to be fused are clustered and information from the clusters is used to inform the fusion method. As we will see below, in the case of microblog search the contribution of information derived from the clusters is very limited. We hypothesize that this is due to the limited length of posts. Therefore, we propose a document expansion technique and hypothesize that it has a positive impact on the performance of cluster-based fusion for microblog search.

We integrate a specific form of document expansion based on semantic linking into the cluster-based fusion method. We first identify entities in each post appearing in any component list; we then expand each post using the text of Wikipedia articles that the entities link to through semantic linking [6, 7]. We do not utilize the document expansion method proposed by [1], as we lack their explicit information generated from additional datasets. Subsequently, we cluster the posts based both on their content and on the additional information from the Wikipedia articles that the entities link to. We then apply cluster-based fusion, which rewards documents that are (possibly) ranked low by the standard fusion method but that are contained in a cluster where many relevant documents are ranked high in many of the lists. To the best of our knowledge, we are the first to integrate semantic linking with fusion for microblog search.

2 Combining semantic linking based cluster-based fusion

Our fusion methods consists of two steps. The first is semantic linking where each post appearing in the lists to be fused is linked to Wikipedia pages, and the second is cluster-based data fusion proper where we create clusters based on the expanded posts.

2.1 Semantic linking We call a sentence in a microblog post a *chunk* so as to be consistent with the literature on semantic linking. In microblog posts, chunks form a sequence $M = \langle m_1, \cdots \rangle$ and our first task is to decide whether a link to Wikipedia should be created for m_i and what the link target should be. A *link candidate* $c_i \in C$ links an *anchor* a in chunk m_i to a target w; an anchor is n-gram in a chunk. Each target w is a Wikipedia article and a target is identified by its unique title in Wikipedia.

In the first step of the semantic linking process, we aim at identifying as many link candidates as possible. We perform lexical matching of each n-gram anchor a of chunk m_i with anchor texts found in Wikipedia, resulting in a set of link candidates C for each chunk m_i that each links to a Wikipedia article w. In the second step, we employ the so-called CMNS method [6] and rank the link candidates in C by considering the prior probability that anchor text a links to Wikipedia article w:

$$CMNS(a, w) = \frac{|C_{a,w}|}{\sum_{w' \in W} |C_{a,w'}|},$$

where $C_{a,w}$ is the set of all links with anchor text a and target w. The intuition is that link candidates with anchors that always link to the same target are more likely to be a correct representation. In the third step, we utilize learning to rerank strategy to enhance the precision of correct link candidates. We extract a set of 29 features proposed in [6, 7], and use a decision tree based approach to rerank the link candidates. Then we obtain three Wikipedia articles that the first three top ranked link candidates link to, and extract the most central sentences from these Wikipedia articles and append them to the microblog post. We call this process *semantic document expansion*.

2.2 Combining clusters and semantic linking We use "SemFuse" to refer to the integration of semantic document expansion with cluster-based fusion. Let q be the underlying information need expressed as a query, d a microblog post (also called a document), s_d the content of the semantic document expansion for d, L a set of component lists to be fused, $\mathcal{C}_L$ be the set of posts that appear in any of the lists, and $F_X(d; q)$ the fusion score for post d computed by a data fusion method X such as CombSUM.

We aim at enhancing the ranking effectiveness of the standard fusion method X for microblog search. For each post in the lists, $d \in \mathcal{C}_L$, we compute the fusion score $F_{SemFuse}(d;q)$ as:

$$F_{SemFuse}(d;q) := (1-\alpha)p(d|q) + \alpha \textstyle\sum_{c \in Cl(\mathcal{C}_L)} p(c|q)p(d, s_d|c), \qquad (1)$$

where $Cl(\mathcal{C}_L)$ is a set of clusters generated by a simple nearest neighbors based approach that utilizes the content of both the posts and the semantic document expansion, c is a cluster, and $p(d|q)$, $p(c|q)$ and $p(d, s_d|c)$ are the probabilities of d being relevant to q, c being relevant to q and both the post and its semantic expansion content being relevant to c, respectively. To estimate $p(d|q)$ in (1), we use Bayes' theorem so that $p(d|q) = \frac{p(q|d)p(d)}{p(q)}$; let $p(q|d) \propto F_X(d;q)$, $p(q) = \sum_{d' \in \mathcal{C}_L} p(q|d')p(d')$ and assume a uniform prior for all documents in $\mathcal{C}_L$ so that we obtain:

$$p(d|q) := \frac{F_X(d;q)}{\sum_{d' \in \mathcal{C}_L} F_X(d';q)}. \qquad (2)$$

To estimate $p(c|q)$, we rewrite it as $p(c|q) = p(q|c) \cdot (\sum_{c' \in Cl(\mathcal{C}_L)} p(q|c'))^{-1}$, where we use a product-based representation and compute $p(q|c)$ as $p(q|c) = \prod_{d \in c} F_X(d;q)^{\frac{1}{|c|}}$, where $|c|$ is the number of documents in c. Note that here the clusters are obtained by utilizing both microblog posts and the corresponding semantic expansion content. Then we obtain our estimation as:

$$p(c|q) := \frac{\prod_{d \in c} F_X(d;q)^{\frac{1}{|c|}}}{\sum_{c' \in Cl(\mathcal{C}_L)} \prod_{d' \in c'} F_X(d';q)^{\frac{1}{|c'|}}}. \qquad (3)$$

Similar to the above estimations, we can conveniently get $p(d, s_d|c)$ as:

$$p(d, s_d|c) := \frac{\sum_{d' \in c} sim(d', s_{d'}||d, s_d)}{\sum_{d'' \in \mathcal{C}_L} \sum_{d' \in c} sim(d', s_{d'}||d'', s_{d''})}, \qquad (4)$$

where $sim(d', s_{d'}||d, s_d)$ is the similarity score between the combination of d' and the semantic expansion $s_{d'}$ and the combination of d and the expansion s_d. We compute this similarity score using symmetric Kullback-Leibler divergence as:

$$sim(d', s_{d'}||d, s_d) := \sum_{t \in d', s_{d'}} p(t|\theta_{d', s'_d}) \log \frac{p(t|\theta_{d', s'_d})}{p(t|\theta_{d, s_d})} + \sum_{t \in d, s_d} p(t|\theta_{d, s_d}) \log \frac{p(t|\theta_{d, s_d})}{p(t|\theta_{d', s'_d})},$$

where t is a token in d or s_d, and $p(t|\theta_{d,s_d})$ is the probability of t given a Dirichlet language model for d and s_d.

After replacing $p(d|q)$, $p(c|q)$ and $p(d, s_d|c)$ by (2), (3) and (4), respectively, in (1), we compute the final SemFuse score and rank documents by $F_{SemFuse}(d;q)$.

3 Experimental setup

To measure the effectiveness of our fusion approach, we work with the Tweets2011 collection [5].[1] The task studied at the TREC 2011 Microblog track was: given a query

[1] http://trec.nist.gov/.

with a timestamp, retrieve at least 30 relevant and interesting tweets. In total, 49 test queries were created and 59 groups participated in the TREC 2011 Microblog track, with each team submitting at most four runs, which resulted in 184 runs[2] [5]. The official evaluation metric was precision at 30 (p@30) [5]. The p@30 scores of these 184 runs varied dramatically, with the best run achieving a p@30 score of 0.4551.

In our experiments below, we sample 12 ranked lists based on their p@30 distribution: 6 runs with all of their p@30 scores over 0.40 (Class 1), and another set of 6 runs with p@30 scores between 0.30 and 0.40 (Class 2). The component runs in Class 1 are clarity1, waterlooa3*, FASILKOM02, isiFDL*, DFReeKLIM30* and PRISrun1. The components runs in Class 2 are KAUSTRerank*, ciirRun1, gut, dutirMixFb, normal and UDMicroIDF.[3] Note that in our experiments, the runs in Class 1 are actually the best 6 ones produced by state-of-the-art microblog search algorithms in TREC 2011 Microblog track, and the name of the run marked with a star symbol indicates that this run utilizes expansion information to search posts. In every class, we use run1, run2, run3, run4, run5 and run6 to refer to the runs in descending order of MAP.

We report the performance with the official TREC Microblog 2011 metric, i.e., p@30, plus p@5, 10, 15, and MAP. Statistical significance of observed differences is tested using a two-tailed paired t-test and is denoted using $^\blacktriangle$ (or $^\blacktriangledown$) for significant differences for $\alpha = .01$, or $^\vartriangle$ (and $^\triangledown$) for $\alpha = .05$. We make comparisons with recent microblog search algorithms (the runs to be fused), standard fusion methods (CombSUM and CombMNZ) as well as state-of-the-art cluster-based fusion methods (ClustFuseCombSUM and ClustFuseCombMNZ [3]), and λ-Merge [9].

4 Results and analysis

We report our experimental results of SemFuse and the 5 baselines in Table 1. The performance of our SemFuse and other 5 baselines fusion methods can beat that of the best result list used in the fusion process (run1) in both classes and on most metrics. Many of these improvements are statistically significant. Particularly, in terms of fusing the lists produced by the best individual microblog search algorithms (Class 1), all of the p@30 scores generated by any of the data fusion method are higher than that of the best record in TREC 2011 Microblog track (0.4551), especially for SemFuse, which obtains a score of 0.5259.

It is clear from Table 1 that ClustFuseX (ClustFuseCombSUM or ClustFuseCombMNZ) cannot beat the standard fusion method (CombSUM and CombMNZ) it integrates in almost all cases, and the performance differences between the two are usually not significant. It is instructive to consider Fig. 1(a), where the optimal value of α in (1) is plotted for ClustFuseCombSUM (black, dotted) and SemFuse (blue, solid); a low optimal value of α means that very little cluster-based information is used in the fusion process, a high value indicates that cluster-based information made a big contribution to the overall performance. We believe that the low optimal value of α for ClustFuseCombSUM is due to the fact that obtaining reasonably coherent clusters of microblog posts is challenging, due to the limited length of posts.

[2] The runs can be downloaded from http://trec.nist.gov/.

[3] Again, details of the runs can be found at http://trec.nist.gov/.

Table 1. Performance on the 12 sample lists. Boldface marks the best result per column; a statistically significant difference between SemFuse and the best baseline fusion method is marked in the upper right hand corner of the SemFuse score. A significant difference with run1 for each fusion method is marked in the upper left hand corner using the same symbols. None of the differences between the cluster-based method and the standard method it incorporates are significant.

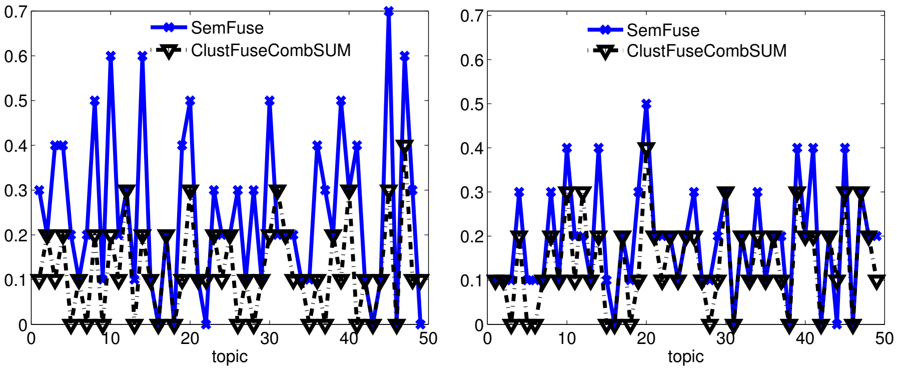
	Class 1					Class 2				
	MAP	p@5	p@10	p@15	p@30	MAP	p@5	p@10	p@15	p@30
run1	.2590	.5959	.5796	.5442	.4537	.1886	.4776	.4347	.3878	.3463
run2	.2575	.5673	.4980	.4721	.4211	.1820	.4122	.3796	.3619	.3027
run3	.2318	.5755	.5367	.5034	.4401	.1688	.3878	.3633	.3605	.3136
run4	.2210	.5918	.5673	.5347	.4551	.1525	.4041	.4143	.3878	.3408
run5	.2098	.5469	.5102	.4694	.4095	.1457	.4612	.4143	.3714	.3571
run6	.2058	.5714	.5367	.4939	.4211	.1376	.3959	.3939	.3796	.3218
CombSUM	$^{\blacktriangle}$.2659	$^{\blacktriangle}$.6245	.5816	.5524	$^{\blacktriangle}$.4966	$^{\blacktriangle}$.1996	$^{\blacktriangle}$.5306	$^{\blacktriangle}$.4531	$^{\blacktriangle}$.4286	$^{\blacktriangle}$.3735
ClustFuseCombSUM	$^{\blacktriangle}$.2655	$^{\blacktriangle}$.6240	.5802	.5503	$^{\blacktriangle}$.4899	$^{\blacktriangle}$.1983	$^{\blacktriangle}$.5287	$^{\triangle}$.4500	$^{\blacktriangle}$.4213	$^{\blacktriangle}$.3686
CombMNZ	$^{\blacktriangle}$.2655	$^{\blacktriangle}$.6245	.5755	.5524	$^{\blacktriangle}$.5020	$^{\blacktriangle}$.1963	$^{\blacktriangle}$**.5347**	$^{\blacktriangle}$.4592	$^{\blacktriangle}$.4354	$^{\blacktriangle}$.3789
ClustFuseCombMNZ	$^{\blacktriangle}$.2650	$^{\blacktriangle}$.6231	.5748	.5502	$^{\blacktriangle}$.4987	$^{\blacktriangle}$.1956	$^{\blacktriangle}$.5330	$^{\blacktriangle}$.4523	$^{\blacktriangle}$.4311	$^{\blacktriangle}$.3731
λ-Merge	$^{\triangle}$.2548	.5641	$^{\triangledown}$.5631	.5496	$^{\triangle}$.4611	$^{\blacktriangle}$.1898	$^{\blacktriangle}$.4641	$^{\blacktriangle}$**.4608**	$^{\blacktriangle}$.4307	$^{\blacktriangle}$.3668
SemFuse	$^{\blacktriangle}$**.2822**$^{\blacktriangle}$	$^{\blacktriangle}$**.6367**$^{\blacktriangle}$	$^{\blacktriangle}$**.5939**$^{\triangle}$	$^{\triangle}$**.5701**$^{\blacktriangle}$	$^{\blacktriangle}$**.5259**$^{\blacktriangle}$	$^{\blacktriangle}$**.2122**$^{\blacktriangle}$	$^{\blacktriangle}$.5306	$^{\blacktriangle}$.4531$^{\blacktriangle}$	$^{\blacktriangle}$**.4435**	$^{\blacktriangle}$**.4000**$^{\blacktriangle}$

SemFuse outperforms all baseline fusions method on class 1, on all metrics, and most of the differences are substantial and statistically significant. As shown in Table 1, the performance of λ-Merge is usually a little lower than that of SemFuse, CombSUM, CombMNZ and the ClustFuseX methods when fusing the lists in Class 1 and Class 2 on almost all metrics. This may be due to its overfitting. The results for the runs in Class 2 are not as clear-cut: the higher the quality of the component result lists, the more improvements can be observed for SemFuse in Table 1. For instance, the p@30 scores after fusion are highest in Class 1 compared to those in Class 2, and the quality of Class 1 is better (p@30>0.4) than that of Class 2 (0.3<p@30<0.4).

Finally, we re-consider the distributions of the free parameter α in (1) that governs the optimal weight of cluster information integrated by SemFuse and ClustFuseX (due to space constraints, we only take ClustFuseCombSUM as a representative) in Figure 1. SemFuse tends to use a bigger optimal weight than cluster-based fusion. This shows that cluster information can effectively be used to boost fusion performance when additional document information is available for generating the clusters; clearly, our semantic document expansion is helpful for cluster-based fusion methods. Another observation from Figure 1 is that the weight differences between SemFuse and ClustFuseCombSUM are more obvious in Class 1 that those in Class 2. We believe that this is due the higher quality of the component lists in Class 1.

5 Conclusion

Microblog search is a challenging IR task because of the special nature of microblog posts. We combine semantic linking with a cluster-based fusion method for searching microblog posts. Our combined fusion method, SemFuse, works with result lists generated by some microblog search algorithms, identifies semantic entities for each post appearing in any of the lists to be fused, appends the most central sentences from the Wikipedia articles that the entities link to to the tweet, and then utilizes the clusters generated from the microblog expansion to enhance cluster-based fusion performance.



(a) Class 1

(b) Class 2

Fig. 1. Distributions of the weight of parameter α on TREC Microblog track 2011 topics when fusing lists in class 1 (left) and class 2 (right) by SemFuse and ClustFuseCombSUM.

Our experiments show that data fusion can improve microblog search performance and semantic document expansion can help to enhance cluster-based fusion methods.

Acknowledgments. This research was supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement nr 288024 (LiMoSINe project), the Netherlands Organisation for Scientific Research (NWO) under project nrs 640.004.802, 727.-011.005, 612.001.116, HOR-11-10, the Center for Creation, Content and Technology (CCCT), the QuaMerdes project funded by the CLARIN-nl program, the TROVe project funded by the CLARIAH program, the Dutch national program COMMIT, the ESF Research Network Program ELIAS, the Elite Network Shifts project funded by the Royal Dutch Academy of Sciences (KNAW), the Netherlands eScience Center under project number 027.012.105 and the Yahoo! Faculty Research and Engagement Program.

6 References

[1] M. Efron, P. Organisciak, and K. Fenlon. Improving retrieval of short texts through document expansion. In *SIGIR '12*, pages 911–920. ACM, 2012.

[2] M. Farah and D. Vanderpooten. An outranking approach for rank aggregation in information retrieval. In *SIGIR '07*, pages 591–598. ACM, 2007.

[3] A. K. Kozorovitzky and O. Kurland. Cluster-based fusion of retrieved lists. In *SIGIR*, pages 893–902, 2011.

[4] S. Liang, M. de Rijke, and M. Tsagkias. Late data fusion for microblog search. In *ECIR '13*, pages 743–746, 2013.

[5] J. Lin, C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC 2011 Microblog track. In *TREC 2011*. NIST, 2011.

[6] E. Meij, W. Weerkamp, and M. de Rijke. Adding semantics to microblog posts. In *WSDM '12*, pages 563–572. ACM, 2012.

[7] D. Odijk, E. Meij, and M. de Rijke. Feeding the second screen: Semantic linking based on subtitles. In *OAIR' 13*, pages 9–16, 2013.

[8] J. A. Shaw and E. A. Fox. Combination of multiple searches. In *TREC 1992*, pages 243–252. NIST, 1993.

[9] D. Sheldon, M. Shokouhi, M. Szummer, and N. Craswell. LambdaMerge: merging the results of query reformulations. In *WSDM '11*, pages 795–804, 2011.