JointSem: Combining Query Entity Linking and Entity based Document Ranking

Chenyan Xiong, Zhengzhong Liu, Jamie Callan, Eduard Hovy
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
cx,liu,callan,hovy@cs.cmu.edu

ABSTRACT

Entity-based ranking systems often employ entity linking systems to align entities to query and documents. Previously, entity linking systems were not designed specifically for search engines and were mostly used as a preprocessing step. This work presents `JointSem`, a joint semantic ranking system that combines query entity linking and entity-based document ranking. In `JointSem`, the spotting and linking signals are used to describe the importance of candidate entities in the query, and the linked entities are utilized to provide additional ranking features for the documents. The linking signals and the ranking signals are combined by a joint learning-to-rank model, and the whole system is fully optimized towards end-to-end ranking performance. Experiments on TREC Web Track datasets demonstrate the effectiveness of joint learning of entity linking and entity-based ranking.

KEYWORDS

Entity Linking, Document Ranking, Entity-based Search

1 INTRODUCTION

Entities and their semantics in knowledge graphs have been used in various aspects of document ranking, for example, query expansion [2, 8], learning-to-rank [7], and text representations [6]. A key step in these entity-based ranking systems is entity linking [1], which aligns the texts in the query and document to the knowledge graph's semantics, and introduces additional ranking evidence for entity-based ranking systems [2, 5–7].

Though significant progress has been made, entity linking and entity-based ranking research were developed separately. Entity linking systems are mostly optimized for their own metrics, which may not suit the needs of entity-based ranking. For example, entity linking systems may prefer high accuracy on several named entity categories [1], while entity-based search systems need high recall on general domain entities to ensure coverage of the query traffic [9]. On the other hand, entity-based ranking systems merely treat the entity linker as a pre-processing step, and use the annotation as a black box. Even the state-of-the-art automatic entity linking systems still make mistakes, especially on short queries [6]. Frequently the noise introduced by the entity linker is the main error source of entity-based ranking systems [4, 7]. Without access to the detailed linking process, the best an entity-based search system can do is manual correction [2, 4] and post-pruning [6, 7].

This work presents `JointSem`, a joint semantic ranking method that combines query entity linking and entity-based document ranking. `JointSem` spots surface forms in the query, links multiple candidate entities to each spotted surface form to avoid over committing (soft-alignment), and ranks documents using the linked entities. The spotting and linking evidence widely used in the entity linking literature are incorporated to weight the entities, and standard entity-based ranking features are used to rank documents. The whole system – spotting, linking, and ranking – is trained jointly by a learning-to-rank model using document relevance labels.

Our experiments on TREC Web Track datasets demonstrate that the joint model is more effective for ranking; significant improvements were found over both word-based and entity-based ranking systems. Our analysis further reveals that the advantage of `JointSem` comes from both the soft-alignment that passes more information to the ranking model, and the joint modeling of spotting, linking, and ranking signals.

2 JOINT SEMANTIC RANKING

This section first describes the spotting, linking, and ranking features used in `JointSem`, and then the joint learning-to-rank model.

2.1 Spotting, Linking, and Ranking

`JointSem` first aligns entities from the knowledge graph to the given query q in a two-step approach: spotting and linking. The spotting step detects surface forms (entity mentions) $S = \{s_1, ...s_i..., s_M\}$ that appear in the query. The linking step aligns each surface form s_i to some candidate entities: $E^i = \{e^i_1, ...e^i_j..., e^i_N\}$. `JointSem` uses a soft-alignment so that multiple candidate entities are kept. Typical spotting and linking features are extracted for the final ranking model to decide the importance of each soft-aligned entity.

Spotting: The spotting step is conducted by looking up the n-grams in the query in a surface form dictionary. The surface form dictionary contains all the possible surface forms (names and aliases) of entities, and is collected from a training corpus, for example, Wikipedia. Following prior convention [1], we start from the *first* word, spot the longest surface forms, and move to the word after the surface form. No overlapped spots are allowed.

In spotting, the following features, $\phi_s(s_i)$, are extracted to describe the reliability of the surface form:

Linked Probability is the probability of a surface form being linked to any entity in the training corpus.

CIKM'17, November 6–10, 2017, Singapore.

ISBN 978-1-4503-4918-5/17/11...$15.00
DOI: https://doi.org/10.1145/3132847.3133048

Table 1: Spotting, linking, and ranking features. Surface Form Features are extracted for each spotted surface form. Entity Features are extracted for each candidate entity from each spot. Entity-Document Ranking Features are extracted for each entity-document pair. The number in brackets is the dimension of the corresponding feature group.

Surface Form Features (ϕ_s)	Entity Features (ϕ_e)	Entity-Document Ranking Features (ϕ_r)
(1) Linked Probability	(1) Commonness	(16) BM25, Coordinate Match, TFIDF and language model with Dirichlet smoothing from entity's textual fields (name and description) to document's fields (title and body)
(1) Surface Form Entropy	(2) Max and Mean Similarity with Query Words	
(1) Top Candidate Entities Margin		
(1) Surface Form Length and Coverage	(2) Max and Mean Similarity with Other Query Entities	
(1) Surface Form Coverage		

Surface Form Entropy is the entropy of the probabilities of a surface form being linked to different entities in the training corpus [6].

Top Candidate Margin is the difference between the probabilities of the surface form being linked to the most frequent entity and the second one.

Surface Form Length and Coverage are the number of words the surface form contains and the fraction of the query words it covers.

Linking: The linking step aligns entities to each spotted surface form. The surface form dictionary contains the mapping from surface forms to its candidate entities, collected from the training corpus. If a surface form has multiple possible candidate entities, *all* of them are linked (soft-alignment).

The relevance of an entity (e) to the query is described by the following linking features $\phi_e(e)$.

Commonness is the probability of the surface form being linked to the entity among all its appearances in the training corpus [6].

Similarity with Query Words: The similarity between a query word and a candidate entity is calculated by the cosine of their pre-trained embeddings (see §3). The max and mean of the entity's embedding similarity to all query words are used as features.

Similarity with Other Query Entities: The similarities between an entity and the top entity (the one with the highest Commonness) of the other spots are calculated using the pre-trained embeddings. The max and mean of these similarity scores are used as its features.

Ranking: The aligned query entities provide many new ranking features. For each linked entity, `JointSem` uses its textual fields as pseudo queries, and extracts entity-document ranking features using standard retrieval models. The ranking features ϕ_r used in this work are: BM25, TFIDF, Coordinate Match, and language model with Dirichlet smoothing, applied on the entity's name and description and the document's title and body [4, 6, 7]. The descriptions are lower-cased and the standard INQUERY stopwords are removed.

The full list of features is shown in Table 1. In a re-ranking setting, all these features are efficient enough to be extracted online if the surface form dictionary, embeddings, and the entity's textual fields are maintained in the memory.

2.2 Joint Learning to Rank

`JointSem` ranks the candidate document d for q using the entity-based ranking features $\phi_r(E, d)$. Additionally, `JointSem` aims to learn how to better utilize the entity-based ranking signals using the surface form features $\phi_s(S)$ and entity features $\phi_e(E)$. The joint ranking model contains three components.

The first part learns the importance of the surface form s_i:

$$f_s(s_i) = w_s^T \phi_s(s_i).$$

The second part learns the importance of the aligned entity e_j^i:

$$f_e(e_j^i) = w_e^T \phi_e(e_j^i).$$

The third part learns the ranking of document for the entity e_j^i:

$$f_r(e_j^i, d) = w_r^T \phi_r(e_j^i, d).$$

The three parts are combined for the final ranking score:

$$f(q, d|\theta, S, E) = \sum_{i=1}^{M} \sum_{j=1}^{N} f_s(s_i) \cdot f_e(e_j^i) \cdot f_r(e_j^i, d), \tag{1}$$

where M is the number of surface forms in the query, N is the number of candidate entities aligned per surface form ($N > 1$), $f(q, d|\theta, S, E)$ is the final ranking score produced by `JointSem`, and $\theta = \{w_s, w_e, w_r\}$ are the parameters to learn.

Training uses standard pairwise learning-to-rank with hinge loss:

$$\theta^* = \text{argmin}_\theta \sum_q \sum_{d^+, d^- \in D_q^{+,-}} [1 - f(q, d^+|S, E) + f(q, d^-|S, E)]_+.$$

$D_q^{+,-}$ is the pairwise document preferences ($d^+ > d^-$). The whole model is differentiable and is optimized by standard back-propagation.

In Equation 1, $f_s(s_i)$ and $f_e(e_j^i)$ together produce the weight, or attention, for the aligned entity e_j^i, which is used to weight the document ranking score $f_r(e_j^i, d)$ produced by the entity. As the whole model is trained jointly by learning-to-rank, the query entity linking is optimized together with the entity-based ranking for better end-to-end ranking performance.

3 EXPERIMENT METHODOLOGY

Dataset: Our experiments use two ClueWeb Category B corpora, TREC Web Track queries and corresponding relevance judgments. ClueWeb09-B has 200 queries from TREC Web Track 2009-2012; ClueWeb12-B13 has 100 queries from TREC Web Track 2013-2014.

All our methods re-rank the top 100 candidate documents from a base retrieval model. On ClueWeb09, the base retrieval is the SDM runs from the well-tuned and widely used EQFE [2]. On ClueWeb12, not all rankings are publicly available from EQFE, so the base retrieval is Indri's default language model, with KSteming, INQUERY stopword removal, and no spam filtering.

Table 2: Overall accuracies of `JointSem` and baselines. Relative performances compared with `LeToR-Qe-Dw` are shown as percentages. Win/Tie/Loss are the number of queries a method improves, does not change, or hurts, compared with `LeToR-Qe-Dw` on NDCG@20. Best results in each metric are marked bold. †, ‡, §, and ¶ indicate statistically significant improvements over `Coordinate Ascent`†, `EQFE`‡, `EsdRank`§, and `LeToR-Qe-Dw`¶, respectively.

	ClueWeb09-B					ClueWeb12-B13				
Method	NDCG@20		ERR@20		W/T/L	NDCG@20		ERR@20		W/T/L
`Lm`	0.1757	−35.63%	0.1195	−34.48%	70/25/99	0.1060	−4.45%	0.0863	−7.09%	40/20/40
`SDM`	0.2496	−8.54%	0.1387	−23.96%	84/28/82	0.1083	−2.41%	0.0905	−2.52%	43/20/37
`RankSVM`	0.2635	−3.46%	0.1544	−15.32%	90/29/75	0.1205	+8.61%	0.0924	−0.45%	39/22/39
`Coordinate Ascent`	0.2681	−1.77%	0.1617	−11.32%	91/28/75	0.1206	+8.70%	0.0947	+1.96%	44/23/33
`EQFE`	0.2448	−10.32%	0.1419	−22.18%	76/28/90	n/a	–	n/a	–	–/–/–
`EsdRank`	0.2644	−3.14%	0.1756	−3.73%	93/25/76	n/a	–	n/a	–	–/–/–
`LeToR-Qe-Dw`	0.2729	–	0.1824	–	–/–/–	0.1110	–	0.0928	–	–/–/–
`JointSem`	**0.3054**†‡§¶	+11.89%	**0.1926**†‡	+5.63%	99/30/65	**0.1314**¶	+18.46%	**0.1076**	+15.93%	54/20/26

When extracting ranking features, the document's title and body are parsed by Boilerpipe with the 'KeepEverythingExtractor'; all parameters of the retrieval models used are kept default.

Baselines: Word-based baselines are unsupervised language model (`Lm`) and `SDM`, and supervised learning-to-rank models, including `RankSVM` and `Coordinate Ascent`. The learning-to-rank baselines were obtained from prior work [6], which used similar experimental conditions.

Entity-based ranking baselines include the widely used EQFE [2] and EsdRank [7]. The ClueWeb09 ranking results were obtained from their authors' websites. Not all of their ClueWeb12 ranking results are publicly available so some comparisons are provided only for ClueWeb09. We also compare with `LeToR-Qe-Dw` with query entities from an off-the-shelf entity linker [3] and similar entity-based ranking features [6]. The main goal of this experiment is to show the effectiveness of joint query entity linking and entity-based ranking; other entity-based ranking systems that use manual annotations or involve document entities are not fair comparisons.

Evaluation Metrics: The TREC Web Track's official evaluation metrics, NDCG@20 and ERR@20, are used. Statistical significance is tested by the permutation test with $p < 0.05$.

Implementation Details: All supervised methods implemented by us are evaluated using the same 10-fold cross validation, done separately on ClueWeb09 or ClueWeb12 queries. In each fold, the training of `JointSem` and its variants were repeated 20 times, and the one with the best *training* loss is used in testing.

The knowledge graph used is Freebase. The surface form dictionary, including the surface forms, their candidate entities, and corresponding commonness scores are obtained from Google's FACC1 annotation on the two ClueWeb corpora. The linked probabilities of the surface forms are calculated on a recent Wikipedia dump (20170420). The word and entity embeddings are trained with the skip-gram model in Google's word2vec toolkit, with 300 dimensions. The training corpus of embeddings is the Wikipedia dump mixed with its duplicate on which the manual entity annotations are replaced by their Freebase Ids. The base retrieval score is added as a ranking feature to `JointSem`.

The training uses batch training and ndam optimization. The maximum entities allowed per surface form (N) is 5; candidate entities not in the top 5 are extremely rare or noise.

4 EVALUATION

This section presents the overall evaluation results and the analysis of `JointSem`'s source of effectiveness.

4.1 Overall Performance

The overall evaluation results in Table 2 demonstrate that jointly modeling the linking of entities and entity-based ranking helps. `JointSem` outperforms all entity-based baselines which also use query annotations and similar ranking evidence, but treat the entity linking step as a fixed pre-processing step. The performances of entity-based systems are also correlated with the entity linking difficulties on corresponding queries. On ClueWeb09 where the entity linking systems perform better [6], directly using TagMe's results is already helpful (`LeToR-Qe-Dw`), and the improvement of joint semantic ranking is relatively smaller (5 – 10%); on ClueWeb12 where query entity linking is harder [6], fully trusting entity linking's results sometimes even fails to outperform word-based ranking, and `JointSem`'s improvements are bigger (15%).

Factoring in the entity linking influences most of the queries. `JointSem` acts rather differently than baselines, improving about half of the queries. The statistical significances are more frequently observed on ClueWeb09 but less on ClueWeb12, although the relative improvements on the latter are higher. Part of the reason is that ClueWeb12 has fewer queries (100), which also makes the learning of the query level models (spotting and linking) less stable.

`JointSem` differs from previous entity-based ranking systems in two aspects: the soft-alignment that introduces multiple entities per spot, and the joint modeling of the spotting, linking and ranking signals towards end-to-end ranking performance. The rest of the experiments study the effectiveness of these two factors.

4.2 Effectiveness of Soft Alignment

This experiment studies the soft-alignment's influence by varying the number of entities allowed per spot in `JointSem`. The results are shown in Figure 1. The y-axis marks the relative improvements compared with `LeToR-Qe-Dw` which uses 'hard alignment' but similar ranking features. The 'k' in `JointSem-Topk` refers to the number of candidate entities considered per spot, selected by their commonness scores.

Table 3: Performances of `JointSem`'s different variations. `JointSem-NoAtt` is the entity-based ranking without attention. `JointSem-SpotAtt` uses the spot attention with entity-based ranking. `JointSem-EntityAtt` uses the entity attention with entity-based ranking. `JointSem-All` is the full model. Relative performances and Win/Tie/Loss are all compared with `JointSem-NoAtt`.

	ClueWeb09-B					ClueWeb12-B13				
Method	**NDCG@20**		**ERR@20**		**W/T/L**	**NDCG@20**		**ERR@20**		**W/T/L**
`JointSem-NoAtt`	0.2919	–	0.1835	–	–/–/–	0.1258	–	0.1012	–	–/–/–
`JointSem-SpotAtt`	0.3005	+2.95%	0.1882	+2.61%	83/55/56	0.1247	−0.88%	0.1010	−0.27%	31/32/37
`JointSem-EntityAtt`	0.2999	+2.74%	0.1872	+2.06%	85/50/59	0.1240	−1.48%	0.1058	+4.52%	36/34/30
`JointSem-All`	**0.3054**	+4.62%	**0.1926**	+5.00%	88/49/57	**0.1314**	+4.44%	**0.1076**	+6.31%	46/27/27

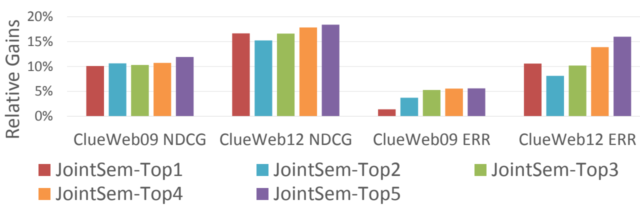

Figure 1: Relative improvements of `JointSem` with different numbers (TopK) of candidate entities per spot. The relative gains marked by the y-axis are compared with `LeToR-Qe-Dw`.

Although in most cases the Top1 entity is the right choice, in general, considering more candidate entities, especially when using all top 5, improves the ranking accuracy. Our manual examination found that improvements are often seen on ambiguous queries whose most popularly linked candidate entities are not the right choice. Without many contexts in short queries, an entity linking system tends to merely choose the most popular candidate entity. For example, the query 'bobcat' refers to bobcat the company, but bobcat the animal is the more popular choice for the entity linking system. `JointSem`'s soft-alignment avoids such over-commitment, and lets the final ranking model select the most useful one(s).

4.3 Effectiveness of Joint Modeling

This experiment studies the effectiveness of joint modeling by comparing `JointSem` and its sub-models. The results are listed in Table 3. `JointSem-NoAtt` uses the Top1 entity per spot as fixed and only includes the ranking part ($f_r(E, d)$). `JointSem-SpotAtt` includes the surface form attention part (f_s), but only the Top1 entities are included with uniform weights; it is similar to the recent attention-based ranking model with word-entity duet [6], but without document entities. `JointSem-EntityAtt` includes the soft-alignment and entity weighting (f_e), but without surface form weighting. `JointSem-All` is the full model. The relative performances and Win/Tie/Loss are all compared with `JointSem-NoAtt`.

The ranking part alone provides better or comparable performance with baselines. Adding in the spotting or the linking part individually helps on ClueWeb09 but has mixed effects on ClueWeb12. Only `JointSem-All` provides stable 5% improvements, confirming the importance of jointly modeling the linking and the utilization of entities for document ranking.

5 CONCLUSION

This work addresses the discrepancy between entity linking and entity-based ranking systems by performing the two tasks jointly. Our method, `JointSem`, spots and links entities in the query, and then uses the linked entities to rank documents. The signals from spotting and linking are incorporated as entity importance features, and the similarities between entities' texts and the document are used as ranking features. `JointSem` uses a joint learning-to-rank model that combines all three components together, and directly optimizes them towards the end-to-end ranking performance.

Experiments on two TREC Web Track datasets demonstrated the effectiveness of `JointSem`, and the influences of the two novelties: the soft-alignment includes multiple entities per spot thus is more robust to ambiguous queries; and the joint modeling stably combines the features from spotting, linking, and ranking together.

This work demonstrates that entity linking, a widely studied natural language processing task, and document ranking, a core information retrieval task, can be, and should be developed together. A future direction is to incorporate entity linking in documents with more advanced entity-based ranking systems.

6 ACKNOWLEDGMENTS

This research was supported by National Science Foundation (NSF) grant IIS-1422676, a fellowship from the Allen Institute for Artificial Intelligence, and DARPA grant FA8750-12-2-0342 under the DEFT program. Any opinions, findings, and conclusions in this paper are the authors' and do not necessarily reflect those of the sponsors.

REFERENCES

[1] David Carmel, Ming-Wei Chang, Evgeniy Gabrilovich, Bo-June (Paul) Hsu, and Kuansan Wang. 2014. ERD'14: Entity recognition and disambiguation challenge. In *Proceedings of SIGIR 2014*. ACM.
[2] Jeffrey Dalton, Laura Dietz, and James Allan. Entity query feature expansion using knowledge base links. In *Proceedings of SIGIR 2014*. ACM.
[3] Paolo Ferragina and Ugo Scaiella. 2010. Fast and accurate annotation of short texts with Wikipedia pages. *arXiv preprint arXiv:1006.3498* (2010).
[4] Xitong Liu and Hui Fang. 2015. Latent entity space: A novel retrieval approach for entity-bearing queries. *Information Retrieval Journal* 18, 6 (2015), 473–503.
[5] Hadas Raviv, Oren Kurland, and David Carmel. 2016. Document retrieval using entity-based language models. In *Proceedings of SIGIR 2016*. ACM, 65–74.
[6] Chenyan Xiong, , Jamie Callan, and Tie-Yan Liu. Word-entity duet representations for document ranking. In *Proceedings of SIGIR 2017*. ACM.
[7] Chenyan Xiong and Jamie Callan. EsdRank: Connecting query and documents through external semi-structured data. In *Proceedings of CIKM 2015*. ACM.
[8] Chenyan Xiong and Jamie Callan. 2015. Query expansion with Freebase. In *Proceedings of ICTIR 2015*. ACM.
[9] Chenyan Xiong, Russell Power, and Jamie Callan. Explicit semantic ranking for academic search via knowledge graph embedding. In *Proceedings of WWW 2017*. ACM.