1 Introduction

Semantic matching in adhoc IR models has been a long desired property. It helps to tackle issues like phrasing differences, out of vocabulary terms, and also provides additional information via soft-matches. Users of IR systems typically expect the system to understand the information need in very few terms. On the other hand, IR systems need more information, for example to tackle disambiguation. Thus intuitively, it is desirable that the model should be able to extract as much information from the query as possible to make better relevance judgements. In this report we look at several models which capture semantic and contextual information in a variety of ways. In particular, we discuss the axioms to govern the behaviour of semantic matches as proposed in [1]. We discuss HiNT [2] which makes use of passage-level signals from whole document and Duet [3] which aims to strike a balance between exact matches and semantic matches. We also consider more recent ELMo [4] and BERT [5] models which produce contextual embeddings. Finally we study CEDR [6], which uses the advances presented in [4] and [5] to improve NIR performance.

2 Semantics Matching

2.1 Fang and Zhai Semantic Matching

It can be argued that synonymy as well as general semantic relatedness is captured by use of mutual information. Occurrences of a terms t_1 and t_2 which are semantically related to (or a synonym of) a term are correlated (e.g. "popular" occurring with "famous" and "well-known"). This translates into higher co-occurrence probabilities used to calculate mutual information. We also note that this method is likely to add noise as terms like "**it**" and "**is**" are highly correlated ("**it is**" co-occurs so often that it has a special contraction "**it's**" in English language). Similarly other pairs of words whose co-occurrence adds no information can have high mutual information. It is interesting to note that the authors propose **TSSC1** as a caution against such false and spurious matches. This is also later reinforced by taking only top K ranked terms (step 3 in section 4.5).

Authors also mention the problem of polysemy by commenting that "an ambiguous term can have multiple senses in a large corpus. [...] the semantically related terms found by mutual information could be a mix of terms corresponding to different senses of the original term". To mitigate this they propose the **TSSC2** constraint. Their approach is to use pseudo-relevance feedback from the result set returned by issuing the original query and then expanding the result using terms from top-M documents. This solution however only works for multi-term queries where this kind of disambiguation is possible and thus will not help with single term queries, where diversification of result set would be a better option.

2.2 HiNT

The authors argue that relevance signals could be concentrated in some passages (or even a single passage) or spread out throughout the document body. Hence making strong assumptions about localisation of such signals is problematic. To address this shortcoming HiNT gathers the relevance evidence from fixed-size passages over the whole document body and learns to combine the signals from different granularity levels. The aim is to capture variety of relevance patterns: only passage P_1 is relevant, P_2, P_5, P_13 are relevant, P_3, P_4, P_5 are relevant etc.

In HiNT, the exact matching matrix M_{ij}^{xor} will clearly not distinguish between the different sense of the word and is thus not interesting for our analysis. We focus on M_{ij}^{cos} which encodes the soft and exact matches together. It is clear that synonymous terms will be captured through semantic matching. For polysemy we distinguish between two cases:

(i) In the first case we assume that the the corpus on which embeddings were trained on had a consistent meaning of the query term and thus the embeddings only represent that particular sense and are oblivious to existence of any other usage. In this case the soft matches will only occur between terms that are related to the predominant usage of the term. We remark that the sense in which a word has been used in the document is relatively consistent within a single document (and hence its passage). It has been known that embeddings retain the sense and biases dominant in the corpus on which embeddings were trained on [7]. This however is equivalent to assuming that polysemy does not exist.

(ii) In the second case we relax the strong assumption of the "clean corpus", but allow multiterm queries. For example instead of just a single term like "book" we consider "**book hotels saarbruecken**" and "**university calculus book**" as two queries. In this case, the exact matches will happen to all the documents containing the query term in any sense. However, the soft matches will only happen to the documents containing the disambiguation terms.

Therefore, a query term q which has more than one senses we can disambiguate most of the times through other terms which are related only to one of the two senses, for example, "**bank deposit**" and "**river bank**". In this sense HiNT and [1] are rather similarly limited in their ability to deal with polysemy.

2.3 Duet

Duet is composed of two jointly trained DNNs each focusing on different matching aspects. The *local model* focuses on exact matches whereas the *global model* performs matching in a learned embedding space. The final score is produce by a sum of individual models' scores. As in HiNT the local model in Duet is not interesting for the purpose of semantic matching. Moreover, only one term at a time is compared with whole document making this local network a BoW model.

The global model begins by representing the query and document terms as n-graph frequency vectors. This representation has a dimension $m_d = 2000$. The query and document terms are limited to 10 and 1000 respectively. This results in a matrix Q of dimension 2000×10 and a matrix D of dimension 2000×1000 for document terms. Each column of these matrices represent a term in 2000-d space. After this, in the similar spirit to Conv-KNRM [8] the embeddings are convolved together in sets of 3 to produce a joint representation. Output of this step is a matrix $\tilde{Q}$ and $\tilde{D}$ of dimension 300×8 and 300×998 respectively. In these matrices each column represents a joint three term embedding with a context of left and right terms. The dimension is reduced via max pooling resulting in a 300×1 vector for query and a 300×889 representation for document. The matching is performed in this space.

It is clear that an individual term's embedding will not be preserved in this method. Thus there is no guarantee of capturing synonyms via semantic matches. For each trigram a different representation is created by Duet. However how this contextual representation could be useful to tackle polysemy is not clear. Also worth mentioning is the fact that the initial character n-graph encoding is useful for dealing with out of vocabulary words but is not meant for semantic matching. In Duet the implication is that the nonlinear transformations applied to the initial representations will result in a suitable representation endowed with meaningful semantic matching. This however is not enough grounds to make strong statements on semantic matching capability of the model.

2.4 BERT

BERT combines the idea of contextual embeddings from ELMo [4] and that of self-attention (in the form of Transformer) from [9] to produce context-dependent embeddings. ELMo presented a way to produce contextual embeddings for downstream NLP tasks using RNNs and the transformer layer introduced in [9] provided an efficient way to incorporate context by use of multi-headed self-attention layer instead of using recurrent architectures like LSTM and RNNs . These capabilities allow BERT to deal with long range dependencies efficiently.

BERT's contextual embedding method is capable of identifying the different senses in which a term has been used. The attention mechanism in BERT can influence the current word's embedding by context words. Thus BERT is more robust to the problems of synonymy and polysemy compared to other (admittedly simpler) models considered in this report.

3 When semantic matching is not enough

In this section we mention some cases where semantic matching even with textual context as in BERT will not be enough to improve retrieval.

(i) In some queries words do not provide enough information to reliably satisfy the information need. For example, consider the query "**pandas**". Without additional

information it is impossible to identify which of the senses is meant: someone looking for panda facts/pictures, pandas antivirus, python pandas package documentation etc. In these cases diversifying the retrieval results conditioned on some prior user model (based on user's previous searches etc) is a good idea.

(ii) Another example is the query "**george bush**". There is no way to tell whether the user in interested in "President George Bush Senior" or "President George Bush Junior" - or perhaps some lesser known "George Bush". It should be remarked here that even query expansion and knowledge-base querying to disambiguate entities will not be helpful here as these two entities overlap in many dimensions. More important context would be the time and whether there is an on going event related to one of the two "George Bush" (e.g. Bush Senior's death).

(iii) Some queries are more sensitive to time. For example given the query "**ind vs pak score**" it is reasonable to assume that with high probability the user it looking for a more recent (or perhaps an ongoing) match between the teams. In this case the information need is very specific but temporal dimension is ambiguous. Other than time, location is also an important background signal. For example the retrieval results for query "**taxi service**" or "**lunch discount deals**" would benefit more from information about user's geographical location to filter out irrelevant results which otherwise match textual relevance signals.

(iv) Another relevant point is that the synonyms can also be context dependent. For example someone looking for "**obama family tree**" would not be amused to see results about "**obama family woods**" or "**obama family forest**". For a more striking example, consider the query "**video games with bugs**" [10]. This could have two senses: video games which have programming bugs or video games where the playable character is an insect. In such cases the synonyms for "bugs" is also sense-dependent. In one case it would be "insect" etc. and in the other case it would be "glitch", "errors", "crashes" etc. While it is tempting to treat a match with the synonym of a term as an exact match and for single term queries it would make sense to have this kind of behaviour but doing so with multi term queries will quickly lead to query drift as discussed. Thus we advise to err on the side of caution and not mix exact match with semantic matches.

4 Passage retrieval and axioms

Given a query q and a document collection C with documents d_i, we want to retrieve a ranked list of passages $[p_1, p_2, ..., p_n]$ relevant to the query. Where each p_i refer to any continuous span of text in a document which answer the information request. With this definition it can be seen that passage retrieval will benefit from the same basic set of signals i.e exact matches, semantic matches, query coverage, term dependence, term proximity, length normalization, and TF-IDF. Hence additional axioms specifically for passage retrieval seem unnecessary. There is however an additional consideration as we have to identify the boundaries (start and end point) of relevance. There is an interesting parallel to this problem in Computer Vision where region growing and region

merging techniques [11] are used to segment a region of interest from an uninformative background. These growing and merging methods propose heuristics which govern when regions in an image can be merged to form a single region. It could be applied to passage retrieval by treating short spans of text as candidate regions and merging / growing them until the similarity gain is under some threshold.

5 A study in CEDR

The reported performance of BoW models i.e. CEDR-KNRM and CEDR-DRMM, is better than the position aware CEDR-PACRR model in [6]. We also note that the performance increase from PACRR to CEDR-PACRR is rather small when compared to the increase from KNRM to CEDR-KNRM, and specially DRMM to CEDR-DRMM. This would suggest that the addition of `[CLS]` input benefits BoW models more than it does PACRR. This suggests the following hypothesis: The embeddings from BERT and `[CLS]` output already incorporate some positional information because of positional embeddings being used at pre-training and fine-tuning time. Thus the additional information provided by PACRR's advanced architecture is at best redundant and at worse confusing for the model to take into account.

This hypothesis can be tested by simply running PACRR in BoW mode i.e. with only 1×1 kernels. We refer to this model as CEDR-BoW-PACRR. If this CEDR-BoW-PACRR achieves the same result as full CEDR-PACRR model then the hypothesis would be validated.

We also remark that the reported results are even more counter-intuitive if one believes that exact matches are crucial for IR systems, which under the usual axioms is a reasonable position. DRMM and KNRM both distinguish between exact and soft matches with a special bin or kernel having $\mu = 1.0$. However, we know that BERT's embeddings do not necessarily satisfy $sim(q, q) = 1$ because of contextual attention (cf. Fig 1. section 3.2 in [6]). This would mean that there are zero entries in the matching histogram or pooling kernel corresponding to exact match. Other than PACRR's results, this observation also raises questions on the origin of performance gains for CEDR-DRMM and CEDR-KNRM.

References

[1] H. Fang and C. Zhai. "Semantic term matching in axiomatic approaches to information retrieval". In: *SIGIR* (2006), pp. 115–122.

[2] Yixing Fan et al. "Modeling Diverse Relevance Patterns in Ad-hoc Retrieval". In: *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval (SIGIR '18)* (2018).

[3] Bhaskar Mitra, Fernando Diaz, and Nick Craswell. "Learning to Match Using Local and Distributed Representations of Text for Web Search". In: *WWW'17* (2017).

[4] Matthew E. Peters et al. "Deep contextualized word representations". In: *NAACL'18* (2018).

[5] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *NAACL'19* (2019).

[6] Sean MacAvaney et al. "CEDR: Contextualized Embeddings for Document Ranking". In: *SIGIR'19* (2019).

[7] Tolga Bolukbasi et al. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". In: *30th Conference on Neural Information Processing Systems NIPS'16* (2016).

[8] Zhuyun Dai et al. "Convolutional Neural Networks for Soft-Matching N-Grams in Ad-hoc Search". In: *WSDM* (2018).

[9] Ashish Vaswani et al. "Attention Is All You Need". In: *31st Conference on Neural Information Processing Systems NIPS'17* (2017).

[10] u/mortimermcmirestinks. *r/gaming subreddit*. URL: https://www.reddit.com/r/gaming/comments/cdegzl/i_mean_youre_not_wrong/. (accessed: 15.07.2019).

[11] C. A. Bouman. *Digital Image Processing*. URL: https://engineering.purdue.edu/~bouman/ece637/notes/pdf/Segmentation.pdf. (accessed: 16.07.2019).