

Supervised Semantic Indexing for Ranking Documents

Bing Bai, Jason Weston, Ronan Collobert and David Grangier
NEC Laboratories America, Princeton, NJ
bbai,jasonw,collober,dgrangier@nec-labs.com

Ranking text documents given a query is one of the key tasks in information retrieval. Typical solutions include classical vector space models using weighted word counts and the cosine similarity (TFIDF) with no machine learning at all, or Latent Semantic Indexing (LSI) using *unsupervised* learning to learn a low dimensional space of “latent concepts” via a reconstruction objective. The former assumes independence of words and cannot capture synonymy or polysemy, whilst the latter is still agnostic to the actual task of interest.

We study *supervised* learning of models of the following type (referred to as Supervised Semantic Indexing (SSI)):

$$f(q, d) = q^\top W d = \sum_{i,j=1}^{\mathcal{D}} q_i W_{ij} d_j \quad (1)$$

where $f(q, d)$ is the score between a query term vector q and a given document term vector d , and $W \in \mathbb{R}^{\mathcal{D} \times \mathcal{D}}$ is the weight matrix, where d_i typically is a (normalized) count of word i from the dictionary of size \mathcal{D} . This model can capture synonymy and polysemy as it looks at all possible cross terms, and can be tuned directly for the task of interest. One can increase efficiency and control capacity by constraining W in the following way:

$$W = U^\top V + I. \quad (2)$$

This induces a low dimensional “latent concept” space similar to LSI. However, it differs in several ways: most importantly it is trained with a supervised signal. Further, U and V differ so it does not assume query and target document should be embedded in the same way, and the addition of the identity term means this model automatically learns the tradeoff between using the low dimensional space and a classical vector space model. To train the model, we employ the margin ranking loss [3] and minimize:

$$\sum_{(q,d^+,d^-) \in \mathcal{R}} \max(0, 1 - f(q, d^+) + f(q, d^-)) \quad (3)$$

where d^+/d^- are relevant/irrelevant documents for query q , respectively. We train this using stochastic gradient descent. Although this loss has been used previously for learning to rank [4, 1], SSI is the first approach to learn to rank directly from word features.

We tested our model on English Wikipedia, assuming a document d is relevant to a query document q if there is a link from q to d . 24M links among 1.8M documents are

Table 1: Document-document ranking on Wikipedia.

Algorithm	Rank Loss	Mean Avg. Prec. (MAP)	Precision@10
TFIDF	0.842%	0.432±0.012	0.1933±0.007
QE	0.842%	0.432±0.012	0.1933±0.007
α LSI + $(1 - \alpha)$ TFIDF	0.721%	0.433±0.012	0.193±0.007
SSI: $W = U^T V + I$	0.158%	0.547±0.012	0.239±0.008

Table 2: Cross-lingual Japanese document-English document ranking on Wikipedia.

Algorithm	Rank Loss	MAP	Precision@10
TFIDF _{EngEng} (ATLAS translated queries)	4.83%	0.290±0.008	0.243±0.008
α CL-LSI _{JapEng} + $(1 - \alpha)$ TFIDF _{EngEng} (ATLAS)	3.31%	0.275±0.009	0.212±0.008
α SSI _{JapEng} + $(1 - \alpha)$ TFIDF _{EngEng} (ATLAS)	0.75%	0.493±0.009	0.377±0.009

randomly split into two portions, 70% for training and 30% for testing. Table 1 shows that SSI strongly outperforms existing techniques TFIDF, LSI and Query Expansion (QE).

As the query and target texts are modeled separately, our approach is easily generalized to other retrieval tasks such as cross-language retrieval. We consider a task analogous to the previous one: given a Japanese query document q_{Jap} that is the mate of the English document q_{Eng} , rank the English documents so that the documents linked to q_{Eng} appear above the others. Table 2 shows that SSI yields improvement over both a translation based method (using Fujitsu’s ATLAS¹), and Cross Language LSI (CL-LSI) [2].

References

- [1] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *ICML 2005*, pages 89–96, New York, NY, USA, 2005. ACM Press.
- [2] S.T. Dumais, T.A. Letsche, M.L. Littman, and T.K. Landauer. Automatic cross-language retrieval using latent semantic indexing. In *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, 1997.
- [3] Ralf Herbrich, Thore Graepel, and Klaus Obermayer. *Large margin rank boundaries for ordinal regression*. MIT Press, Cambridge, MA, 2000.
- [4] T. Joachims. Optimizing search engines using clickthrough data. In *ACM SIGKDD*, pages 133–142, 2002.

Topic: semantic indexing, learning to rank
Preference: oral

¹<http://www.fujitsu.com/global/services/software/translation/atlas/>