Modelling Term Dependence with Copulas

Carsten Eickhoff
Dept. of Computer Science
ETH Zurich, Switzerland
ecarsten@inf.ethz.ch

Arjen P. de Vries
CWI Amsterdam
Amsterdam, The Netherlands
arjen@acm.org

Thomas Hofmann
Dept. of Computer Science
ETH Zurich, Switzerland
thomas.hofmann@inf.ethz.ch

ABSTRACT

Many generative language and relevance models assume conditional independence between the likelihood of observing individual terms. This assumption is obviously naïve, but also hard to replace or relax. There are only very few term pairs that actually show significant conditional dependencies while the vast majority of co-located terms has no implications on the document's topical nature or relevance towards a given topic. It is exactly this situation that we capture in a formal framework: A limited number of meaningful dependencies in a system of largely independent observations. Making use of the formal copula framework, we describe the strength of causal dependency in terms of a number of established term co-occurrence metrics. Our experiments based on the well known *ClueWeb'12* corpus and TREC 2013 topics indicate significant performance gains in terms of retrieval performance when we formally account for the dependency structure underlying pieces of natural language text.

Categories and Subject Descriptors

Information Systems [**Information Retrieval**]: Retrieval models

Keywords

Relevance models; Multivariate relevance; Ranking; Probabilistic framework; Language models.

1. INTRODUCTION & RELATED WORK

Generative n-gram language models are frequently used tools for representing document or collection vocabulary in the form of probability distributions over (spans of) textual tokens. They are popular for a wide array of tasks, including sentiment analysis, machine translation, content based classification and document retrieval. Most state-of-the-art models assume individual terms to be independently drawn from the underlying distribution. While this independence assumption greatly simplifies the computation of conditional probabilities, it is also rather naïve. It is easy to find examples of proper names such as "*Barack Obama*" or "*Hong Kong*", but also other fixed expressions ("*tax evasion*") as well as non-consecutive constructs ("*dollar*" , "*stock*"), that should benefit from an explicit model of term interdependency. It is easy to see that some of these examples go beyond the capabilities of mere higher-order language models. In the past, there have been a number of attempts to formally integrate term dependency structures into generative models.

Van Rijsbergen's early work on dependency trees [19] theoretically establishes the use of maximum spanning trees and term co-occurrence statistics in order to establish local term dependency structures. Yu et al. [20] present a comparison of tree and cluster-based methods for dependency modelling in information retrieval. Srikanth and Srihari [18] investigate dependency-aware relevance models by using higher order n-gram models and comparing to the unigram setting. They report consistent improvements for the dependency models. Croft et al. [7] propose the use of inference networks trained on the basis of term proximity information to model the relevance of entire phrases. In a related effort, Lossee [12] confirms the important role of proximity in dependency modelling. The author reports optimal performance using context windows of 3-5 terms surrounding each candidate term. In a comparison study of multiple dependency models, Bruza and Song [4] achieve best results using a matrix representation of co-occurrence contexts to describe terms. Gao et al. [11] explicitly decouple the dependency structure of a sentence from the concrete term generation probabilities in the form of linkages.

Nallapati and Allan [14] relax the independence assumption by modelling documents as groups of (still independent) sentences. Within each sentence, however, they condition the probability of observing a term on all previous terms in the same sentence. Their sentence model is based on the maximum spanning tree over the fully connected sentence graph. The individual strength of dependency within term pairs is measured in terms of the Jaccard coëfficient. They later refine this model by advancing from dedicated sentence trees to entire forests [15] of trees for each connected component in the sentence graph. Cao et al. [5] use Bayesian networks to combine two sources of term dependence: co-occurrence and semantic relatedness. The latter is expressed in terms of proximity in the WordNet graph. While most approaches define document language models and compare their respective likelihoods of having generated the query,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
SIGIR'15, August 09 - 13, 2015, Santiago, Chile.
© 2015 ACM. ISBN 978-1-4503-3621-5/15/08 ...$15.00.
http://dx.doi.org/10.1145/2766462.2767831

Bai et al. [1] propose a term-dependency query model in order to account explicitly for query expansion. Metzler and Croft [13] learn Markov Random Fields that account for various forms of term dependence on the basis of arbitrary feature vectors. Bendersky et al. [3, 2] propose a learning to rank approach that accounts for higher order term and concept dependencies using hypergraphs. Shi and Nie [17] investigate different dependence weighting schemes based on the concept's utility in the respective retrieval task.

In this paper, we present the use of copulas, a robust statistical model family that is able to explicitly decouple marginal observations (i.e., the individual likelihoods of generating terms) from their underlying dependency structures. Comparing to a number of well-known baseline methods and evaluating on a large-scale standard dataset, we show the competitive performance of this novel approach to dependency modelling.

2. METHODOLOGY

In this section, we will give a brief overview of the formal copula framework, introducing the relevant notation and conventions. For a more comprehensive introduction to the topic, please refer to the previous IR applications by Eickhoff et al. [8, 9], or the surveys by Embrechts [10] and Schmidt [16].

Let X be a k-dimensional random vector of observations that we wish to use as input to our copula model:

$$X^k = (x_1, x_2, \dots, x_k)$$

The copula allows us to model the likelihood of observing X by offering computationally efficient approximations to the true joint probability distribution in the high-dimensional space of cardinality k. As a first step, copulas require input scores U^k to be uniformly distributed in the $[0 \dots 1]$ interval. We can achieve this by defining a set of transformations $F(X)$ between raw marginal observations X and their normalized equivalents on the unit cube U.

$$U^k = (u_1, u_2, \dots, u_k) = F(X) = (f_1(u_1), f_2(u_2), \dots, f_k(u_k))$$

An easy example of such a function is the empirical distribution function $\hat{f}$:

$$\hat{f}(t) = \frac{1}{n} \sum 1 \{x_i \leq t\}$$

The cumulative distribution function C for all copulas is fully defined in terms of a generator ψ and its inverse ψ^{-1}:

$$C(u_1, u_2, \dots, u_k) = \psi^{-1}(\psi(u_1) + \psi(u_2) + \dots + \psi(u_k))$$

There are many concrete instantiations of such copula functions. Each copula family defines their own generator and inverse. A previous study [8] compared a wide range of copula families for the task of Web retrieval, finding Gumbel copulas to be the most adequate choice in this setting. For reasons of space, this paper builds on the previous findings and concentrates exclusively on Gumbel copulas. Their generators are given in the following form:

$$\psi^{-1}(t) = exp(-t^{\frac{1}{\theta}})$$
$$\psi(t) = (-log(t))^{\theta}$$

The resulting distribution function for a 2-dimensional Gumbel copula is, for example:

$$C(u_1, u_2) = exp(-((-log(u_1))^{\theta} + (-log(u_2))^{\theta})^{\frac{1}{\theta}})$$

Once the choice of copula family is made, we are left with just a single parameter θ that allows us to control the strength of dependency between the individual marginal observations u. If we, for example, set $\theta = 1$, our distribution function defaults to the case of conditional independence:

$$C_{\theta=1} = exp(-(-log(u_1)) + (-log(u_2))) = u_1 * u_2$$

Any choice of $\theta > 1$ results in an increasing degree of conditional dependency between the k dimensions of our observation. At this point, we have introduced all relevant components for our original use case of statistical language modelling. A traditional unigram language model describes the likelihood of observing a string of text T under a given class c as the product of the individual likelihoods of each term:

$$P(T|c) = P(t_1|c)P(t_2|c) \dots P(t_{|T|}|c)$$

The same can be achieved under the copula framework by considering the class-conditional probabilities of observing individual terms $t_1, t_2, \dots$ as our marginal observations, making the dimensionality of our copula $k = |T|$:

$$P(T|c) = C_{\theta=1}(P(t_1|c), P(t_2|c), \dots P(t_{|T|}|c))$$

By choosing $\theta = 1$, we ensure conditional independence between the marginal term observation likelihoods, giving us the standard unigram language model. As we however increase θ, the strength of dependency between the individual terms increases. This ability to account for term dependence makes the copula framework a powerful alternative to the standard language modelling scheme. At this point, any setting of θ globally describes the relationship between all terms. In practice, however, we much rather want a select few terms to depend on each other, while the majority of terms occur indeed independently.

This is easily achieved by using nested copulas. Instead of combining all dimensions in a single step as described earlier, they allow for a nested hierarchy of multiple copulas that estimate joint distributions for sub sets of the full term space and subsequently combine scores until one global model is obtained. Generally, an example of a fully nested copula with k dimensions is given by:

$$C_0(u_1, C_1(u_2, C_2(\dots, C_{k-2}(u_{k-1}, u_k))))$$

By means of the structure of the nesting "*tree*", nested copulas can explicitly model which dimensions depend on each other directly. Instead of the global θ parameter discussed earlier, each of the constituent copulas defines their respective θ_i, determining the strengths of these (per-dimension) dependencies. This mechanism gives nested copulas a theoretical advantage in flexibility over their non-nested counterparts. Effectively, this allows us to describe formally grounded probabilistic models under which select term pairs show dependencies ($\theta > 1$) while the majority of terms oc-

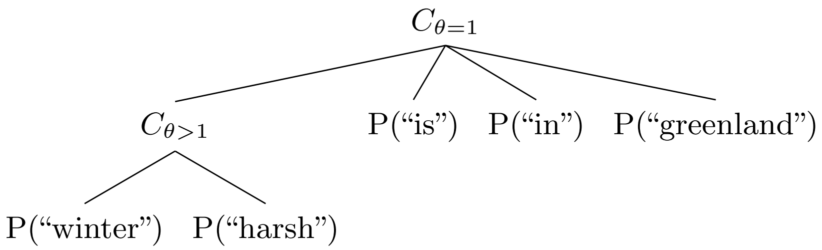

Figure 1: The copula dependence tree for the sentence "*Winter is harsh in Greenland.*" shows significant dependence between the terms "*winter*" and "*harsh*" while the remaining terms occur independently.

cur independently of each other ($\theta = 1$). Figure 1 shows an example of such a situation.

At this point, the final missing component in our language modelling scheme is a way to determine the concrete settings of θ for a pair of terms. To this end, we define conditional dependency in terms of frequency of co-occurrence in a document corpus and rely on two widely used co-occurrence metrics. The point-wise mutual information between terms t_1 and t_2 as well as their Jaccard coefficient measure in which fraction of sentences the terms co-occur.

$$PMI(t_1, t_2) = log_2 \frac{P(t_1,t_2)}{P(t_1)P(t_2)}$$

$$J(t_1, t_2) = \frac{|t_1 \cap t_2|}{|t_1 \cup t_2|}$$

Finally, the dependency parameter θ_{t_1,t_2} is defined on the basis of the concrete choice of metric $m \in \{PMI, J\}$ and its collection-wide metric mean (μ_{PMI}, μ_J) across all potential term pairs. All those pairs of higher-than-average co-occurrence frequency are assigned values of θ proportionally to their relative co-occurrence rate. Since θ is defined in the range $[1, \infty)$ and the resulting scores scale in a non-linear fashion, there is no need to further address or remove outlier pairs of extremely high frequency.

$$\theta_{m,t_1,t_2} = \begin{cases} \frac{m(t_1,t_2)}{\mu_m} & if\ m(t_1, t_2) > \mu_m \\ 1 & else \end{cases}$$

3. EXPERIMENTS

To empirically test the performance of the previously presented copula-based language modelling scheme, we investigate its performance at the task of adhoc document retrieval. Instead of modelling the likelihood of observing a given document under a topic specific language model, we will now establish one distinct model per document and compare their respective likelihoods of having generated the query q.

$$P(rel|q, d) \approx P(q|d)$$

$$P(q|d)_{indep} = \prod_{i=1}^{|q|} P(w_i|d)$$

$$P(q|d)_{cop} = C_d(w_1, w_2, \ldots, w_n)|w \in q$$

For our experimental comparison, we rely on the widely used ClueWeb'12 corpus, a collection of 730 million authentic Web documents. Our 50 topics originate from TREC's 2013 Adhoc retrieval task [6]. We contrast our method's performance with a number of established as well as state-of-the-art baselines such as standard unigram and bigram language models, Nallapati's sentence trees [14], as well as the Markov Random Field model [13] and apply Laplace smoothing to all LM variants in order to account for previously unseen query terms. Table 1 details the respective performances obtained by the various methods in terms of precision, recall, F_1 and MAP, each computed at a cut-off rank of 20 retrieved documents. Statistically significant improvements over all baseline methods are indicated by the asterisk character. Statistical significance was tested using a Wilcoxon signed-rank test at $\alpha \leq 0.05$-level. We can note that, due to their wider context, the classification performance of bigram language models significantly exceeds that of the lower-order model. SenTrees as well as the MRF model which explicitly capture term dependence show even higher classification performance. Finally, our copula language model yields significant performance improvements across most metrics and baselines. The improvements over the MRF model were only significant for some of the considered metrics.

Table 1: Retrieval performance on ClueWeb'12 and TREC 2013 Adhoc topics at a cut-off threshold of 20 retrieved documents.

Model	Precision	Recall	F_1	MAP
Unigram LM	0.31	0.22	0.26	0.41
Bigram LM	0.34	0.26	0.3	0.45
SenTree	0.35	0.28	0.31	0.47
MRF	0.38	0.31	0.34	0.51
Copula LM	**0.41***	**0.35***	**0.38***	**0.52**

4. CONCLUSION

In this paper, we demonstrated the use of the copula framework, a model family from the field of robust statistics, for representing term dependencies in language models. The main advantages of the proposed model are its formal rigour, the low model complexity in terms of training effort as well as disk space requirements and its high degree of flexibility. As an additional advantage, copulas have been previously shown [9] to be beneficial for qualitative manual inspection of results.

Our experiments, based on a sizeable document collection (ClueWeb'12) confirm the competitive performance of the proposed model in comparison with a number of state-of-the-art baselines.

The present paper describes early results of an ongoing body of research. Consequently, there are numerous directions for future work that are interesting to explore: **(1)** In this paper, we investigated "single-layer" dependency structures with a nesting depth of 1. The nested copula framework, however, is able to capture arbitrarily complex structures. Given a modified dependency estimation scheme, the model can easily account for cases of fine-grained multi-level dependencies. **(2)** Similarly, the current model regards only dependencies of degree 2. Since the copula framework is able to account for higher-degree dependencies (i.e., between three or more terms) this is another promising alley for continued research. **(3)** Previous work has investigated different forms of inter-term dependency, including, for example, semantic proximity. It would be easy to integrate such additional sources of evidence into our θ estimation step. **(4)** We would like to draw from the existing wealth of topic modelling techniques in order to describe not merely the dependency structure between individual terms but also between terms and more high-level (latent) concepts, allowing for exciting new insights. **(5)** Finally, we would like to explore means of representing local term context into the dependency model. Take for instance the two terms "new" and "york". In the query [affordable real estate on the US east coast] these terms clearly have some dependency, whereas they probably are less dependent when the query is [affordable real estate in yorkshire, UK]. The current paper highlights the applicability of the context invariant method for general queries. In the future we will additionally investigate the importance of the immediate context.

5. REFERENCES

[1] Jing Bai, Dawei Song, Peter Bruza, Jian-Yun Nie, and Guihong Cao. Query expansion using term relationships in language models for information retrieval. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 688–695. ACM, 2005.

[2] Michael Bendersky and W Bruce Croft. Modeling higher-order term dependencies in information retrieval using query hypergraphs. In *Proceedings of the 35th international ACM SIGIR conference*, pages 941–950. ACM, 2012.

[3] Michael Bendersky, Donald Metzler, and W Bruce Croft. Learning concept importance using a weighted dependence model. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 31–40. ACM, 2010.

[4] Peter Bruza and Dawei Song. A comparison of various approaches for using probabilistic dependencies in language modeling. In *Proceedings of the 26th annual international ACM SIGIR conference*, pages 419–420. ACM, 2003.

[5] Guihong Cao, Jian-Yun Nie, and Jing Bai. Integrating word relationships into language models. In *Proceedings of the 28th annual international ACM SIGIR conference*, pages 298–305. ACM, 2005.

[6] Kevyn Collins-Thompson, Paul Bennett, Fernando Diaz, Charles LA Clarke, and Ellen M Voorhees. Trec 2013 web track overview. In *22nd Text REtrieval Conference, Gaithersburg, Maryland*, 2014.

[7] W Bruce Croft, Howard R Turtle, and David D Lewis. The use of phrases and structured queries in information retrieval. In *Proceedings of the 14th annual international ACM SIGIR conference*, pages 32–45. ACM, 1991.

[8] Carsten Eickhoff and Arjen P de Vries. Modelling complex relevance spaces with copulas. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1831–1834. ACM, 2014.

[9] Carsten Eickhoff, Arjen P de Vries, and Kevyn Collins-Thompson. Copulas for information retrieval. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 663–672. ACM, 2013.

[10] P. Embrechts, F. Lindskog, and A. McNeil. Modelling dependence with copulas and applications to risk management. *Handbook of heavy tailed distributions in finance*, 8(329-384):1, 2003.

[11] Jianfeng Gao, Jian-Yun Nie, Guangyuan Wu, and Guihong Cao. Dependence language model for information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 170–177. ACM, 2004.

[12] Robert M Losee. Term dependence: truncating the bahadur lazarsfeld expansion. *Information processing & management*, 30(2):293–303, 1994.

[13] Donald Metzler and W Bruce Croft. A markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference*, pages 472–479. ACM, 2005.

[14] Ramesh Nallapati and James Allan. Capturing term dependencies using a language model based on sentence trees. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 383–390. ACM, 2002.

[15] Ramesh Nallapati and James Allan. An adaptive local dependency language model: Relaxing the naıve bayes' assumption. 2003.

[16] T. Schmidt. Coping with copulas. *Risk Books: Copulas from Theory to Applications in Finance*, 2007.

[17] Lixin Shi and Jian-Yun Nie. Using various term dependencies according to their utilities. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1493–1496. ACM, 2010.

[18] Munirathnam Srikanth and Rohini Srihari. Incorporating query term dependencies in language models for document retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference*, pages 405–406. ACM, 2003.

[19] Cornelis Joost van Rijsbergen. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of documentation*, 33(2):106–119, 1977.

[20] Clement T Yu, Chris Buckley, K Lam, and Gerard Salton. A generalized term dependence model in information retrieval. Technical report, Cornell University, 1983.