A Comparison of Various Approaches for Using Probabilistic Dependencies in Language Modeling

Peter Bruza and Dawei Song
Distributed Systems Technology Centre
Level 7, General Purpose south
University of Queensland, QLD 4072 Australia

Categories and Subject Descriptors: H.3.3 [**Information Search and Retrieval**]: Retrieval Models – Language models

General Terms: Theory, Algorithms

Keywords: Probabilistic dependencies, Hyperspace analogue to language, Information flow

1. INTRODUCTION

The Relevance-based language model is a promising invention within the language modeling approach to document retrieval [4]. The relevance model computes $\Pr(w \mid R)$ which is interpreted as the "probability of observing a word w in documents relevant to an information need". In practice, this probability is approximated by $\Pr(w \mid q_1, q_2, ..., q_k)$ for a query Q = $(q_1, q_2, ..., q_k)$. This probability can be computed in terms of the joint probability of *w* and Q:

$$\Pr(w \mid q_1, q_2, ..., q_k) = \frac{\Pr(w, q_1, q_2, ..., q_k)}{\Pr(q_1, q_2, ..., q_k)} \qquad (1)$$

The goals of this article is to study several estimates of relevance models which will be computed based on differing approaches for incorporating term dependency information. In this way, we hope to shed light on the relative merits of term dependency information, as well as provide a theoretical framework for such investigations.

2. COMPUTING PROBABILISTIC DEPENDENCIES

In order to incorporate term dependencies into a retrieval model, they must be captured in an expedient way. This section details a probabilistic variant of the Hyperspace Analogue to Language model (HAL) [1].

HAL Spaces

Given an *n*-word vocabulary, the HAL space is constructed by moving a sliding window over the corpus by one word increment. All words within the window are considered as co-occurring with each other with strengths inversely proportional to the distance between them. After traversing the corpus, an *n x n* term matrix which can be used to produce vector representations of words [1][2]. The following is part of a normalized HAL vector for "*superconductors"* computed from the AP corpus:

superconductor = < commercial:0.15 consortium:0.18 develop:0.12 electricity:0.18 energy:0.07 high:0.34 materials:0.25 new:0.24 research:0.12 resistance:0.13 scientists:0.11 semiconductors:0.10 temperature:0.48 ….>

HAL has had notable success in producing vector representations of words with cognitive compatibility of HAL vectors with human processing, for example, word matching and word similarity experiments. HAL has also been used as the basis of an information inference mechanism used for query expansion with encouraging results [2]. In summary, HAL seems to capture useful associations between words, which can be interpreted probabilistically.

Probabilistic HAL Spaces

Row *i* in a HAL matrix represents weighted associations of term *i* to other words seen in the context of *i*, summed across the whole collection. The HAL vector above describing "superconductor" is an example of such a row. By normalizing the weights in a vector, and then squaring them, conditional probabilities result. For example, in the context of the example HAL vector, $\Pr(superconductor|temperature) = 0.48^2 = 0.2304$. If a term *j* has not appeared in the same context as word *i*, then *i* is assumed to be conditionally independent of *j* : $\Pr(i \mid j) = \Pr(i)$.

3. INCORPORATING PROBABILISTIC DEPENDENCIES VIA APPROXIMATION TO CHAIN RULE

Having built a probabilistic HAL space, one way of incorporating conditional dependencies between terms is to employ the Chain rule [3]: $\Pr(w, q_1, \ldots, q_k) =$

$$\Pr(w)\Pr(q_1 \mid w)\Pr(q_2 \mid w, q_1)\ldots\Pr(q_k \mid w, q_1, \ldots, q_{k-1}) \quad \text{(Chain - rule)}$$

The intuition behind this formula is founded on the fact utterances in a language are not random: Given a sequence of words, the next word is dependent on words previous in the sequence.

In practice, the Chain Rule has the formidable problem of requiring reliable estimates of the conditional probabilities comprising the chain. Therefore, various simplifying assumptions are made to approximate the Chain Rule. For notational convenience, these approximations will be detailed in terms of the following rendering of the Chain Rule: $\Pr(q_0, q_1, \ldots, q_k) =$

$$\Pr(q_0)\Pr(q_1 \mid q_0)\Pr(q_2 \mid q_0, q_1)\ldots\Pr(q_k \mid q_0, q_1, \ldots, q_{k-1})$$

First-order Markov approximation of the Chain Rule

By assuming that a given term in the sequence is only dependent on the previous term leads to a first order Markov approximation [6]. More formally, $\Pr(q_i \mid q_0, \ldots, q_{i-1})$, *i>1* is approximated with $\Pr(q_i \mid q_{i-1})$. This is also known as bi-gram language modeling. In terms of the first-order limited horizon, the Chain Rule can be approximated as follows:

$$\Pr(q_0, q_1, \ldots, q_k) = \Pr(q_0)\prod_{1 \le i \le k}\Pr(q_i \mid q_{i-1}) \quad \text{(Markov - 1)}$$

The advantage of the first-order limited horizon is that conditional dependencies only involve single terms q_i, q_j in $\Pr(q_i \mid q_j)$. Conditional probabilities of this form are easier to estimate than that involving multi-term evidence (higher-order dependencies).

Another Markov approximation has been proposed in which a broader horizon is considered, but higher order dependencies are avoided by using the conditional term dependency contributing most within the horizon [7]:

$$\Pr(q_0, q_1, \ldots, q_k) = \Pr(q_0)\prod_{1 \le i \le k}\max_{0 \le j < i}\{\Pr(q_i \mid q_j)\} \quad \text{(Markov - Max)}$$

Conditional Sampling

Conditional sampling assumes the query terms $q_1, \ldots, q_k$ to be independent of each other, but the dependence on term *w* is kept [4].

$$\Pr(w, q_1, \ldots, q_k) = \Pr(w) \prod_{1 \le i \le k} \Pr(q_i \mid w) \quad \text{(CondSamp)}$$

The following approximation of this formula was used with encouraging results on TREC retrieval and tracking tasks [4]. Instead of using the probabilistic HAL space, conditional probabilities $\Pr(q_i \mid w)$ are computed over a universe $\mathbf{M}$ of unigram models. It is additionally assumed that q_i is independent of *w* once the distribution M_i has been chosen.

$$\Pr(w, q_1, \ldots, q_k) = \Pr(w) \prod_{1 \le i \le k} \sum_{M_i \in \mathbf{M}} \Pr(M_i \mid w) \Pr(q_i \mid M_i) \quad \text{(CondSamp - X)}$$

Note that all of the above formulae (Markov-1, Markov-Max, CondSamp, CondSamp-X) can be used to compute $\Pr(w \mid q_1, \ldots, q_k)$ by equation (1), where $\Pr(q_1, \ldots, q_k) = \sum_w \Pr(w, q_1, \ldots, q_k)$.

4. HIGHER ORDER CONDITIONAL PROBABILITIES VIA INFORMATION FLOW

Recent research into HAL-based information inference allows to compute the degree of information flow, denoted degree($q_1 \oplus \ldots \oplus q_k \mid\!- w$), to which a term *w* can be inferred "informationally" from the combination of query terms (denoted $q_1 \oplus \ldots \oplus q_k$). For detailed formula of information flow computation please refer to [9]. The following example shows some top ranked information flows (and their degrees) derived from the term combination "gatt ⊕ talks" (GATT - General Agreement on Tariffs & Trade is a forum for global trade talks).

gatt ⊕ talks |- { gatt: 1.0, trade: 0.96, agreement: 0.96, world: 0.86, negotiations: 0.85, talks: 0.84, set: 0.82, states: 0.82, EC: 0.81, japan: 0.78, farm: 0.78, rules: 0.76, round: 0.76, members: 0.74, council: 0.73, agriculture :0.73, officials: 0.72, government: 0.72, … }

The combination of query terms is important in IR, as combinations of words in a query topic may represent a single underlying concept, e.g., "star wars" etc. The HAL vectors of query terms are combined into a single vector via a heuristic form of vector addition [2].

Information flow can then used to compute the higher order conditional probability directly:

$$\Pr(w \mid q_1, \ldots, q_k) \cong \frac{\text{degree}(q_1 \oplus q_2 \oplus \ldots \oplus q_k \mid\!- w)^2}{\sum_w \text{degree}(q_1 \oplus q_2 \oplus \ldots \oplus q_k \mid\!- w)^2} \quad \text{(InfoFlow)}$$

5. EMPIRICAL COMPARISON OF DEPENDENCY MODELING APPROACHES

We empirically compare the above five methods (i.e., Markov-1, Markov-max, CondSamp, CondSamp-X, and InfoFlow) by measuring the "closeness" between the true relevance model and query modes constructed using these approaches. The experiment uses AP 88&89 collection (164, 597 documents and 249, 453 non-stop words) and TREC topics 101-200. Only titles are used as queries.

The true relevance model is the unigram distribution over the relevant documents (for a given query Q). More formally, $P_R(w) = \frac{tf(w)}{\sum_{v \in V} tf(v)}$, where $tf(w)$ is the term frequency of *w* in the collection of relevant documents and V is the vocabulary of terms from this collection. The relevance model is not smoothed.

Each of the above approaches is then used to calculate an estimate of the relevance model, i.e. Pr(w|Q), for all words over the vocabulary. The Kullback-Leibler divergence is employed to measure the divergence of an estimate P_e of the relevance model and the true model P_R :

$$KL(P_R, P_e) = \sum_w P_R(w) \log \frac{P_R(w)}{P_e(w)}$$

The result of the comparison is shown below. The associated number with each model is its KL divergence from the true relevance model.

Discussion: A relevance-based query model created by InfoFlow is the closest to the true relevance model, followed by Markov-1, Markov-Max, CondSamp and CondSamp-X. From an intuitive point of view, this ordering seems consistent with the level of sensitivity with regard to how dependency information is incorporated in each model: the InfoFlow uses high-order conditional probabilities whereby all query terms are considered as evidence (context) and no independence assumption between query terms is made, the Markov-1 and Markov-Max use bi-gram conditional probabilities, and the two conditional sampling approaches actually assume a relaxed independence assumption (*w* is dependent to the query terms, but query terms are independent of each other). Markov-1ranking in front of Markov-Max may indicate that the maximum valued dependency of a term is not as reliable for probabilistically estimating the context of a term as using the previous term.

6. CONCLUSION AND FUTURE WORK

By ranking estimates of query-based relevance models with the true relevance model using the Kullback-Leibler divergence provides a mechanism for potentially predicting the relative performance of models incorporating probabilistic dependencies in various ways. The next task is to examine whether the predicted ranking actually occurs in practice via traditional recall-precision experiments. To this end, probabilistic HAL spaces will be used to capture the probabilistic dependencies. Document language models will be constructed and ranked with respect to the estimate of the relevance model, for example by using the KL-divergence as in [5].

Acknowledgements

The work reported in this paper has been funded in part by the Cooperative Research Centres Program through the Department of the Prime Minister and Cabinet of Australia.

REFERENCES

[1] Burgess, C., Livesay, K. and Lund K. (1998) Explorations in Context Space: Words, Sentences, Discourse. *Discourse Processes, 25(2&3), 211-257.*

[2] Bruza, P.D., and Song, D. (2002) Inferring Query Models by Computing Information Flow. In Proceedings of CIKM 2002, pp. 260-269.

[3] Charniak, E. (1993) Statistical Language Learning. MIT Press.

[4] Lavrenko, V. and Bruce Croft, W. (2001) Relevance-Based Language Models. In Proceedings of ACM SIGIR'2001, pp. 120-127.

[5] Lavrenko, V., Choquetee, M. and Bruce Croft, W. (2002) Cross-Lingual Relevance Models. In Proceedings of ACM SIGIR'2002, pp. 175-182.

[6] Manning, C.D. and Schütze, H. (2002). *Foundations of Statistical Natural Language Processing.* MIT Press.

[7] Nallapati, R. and Allan, J. (2002) Capturing Term Dependencies using a Language Model based on Sentence Trees. In Proceedings CIKM 2002, pp. 383-390.

[8] Van Rijsbergen, C.J. (1979) *Information Retrieval*. Butterworths

[9] Song, D., and Bruza, P.D. (2003) Towards Context-sensitive Information Inference. *JASIST*, 54(4), pp. 321-334..

[10] Turtle, H.R. and Croft, B. (1990). Inference Networks for Document Retrieval. Proceedings of SIGIR'1990, pp.1-24.