Introduction to the Special Issue on Statistical Language Modeling

JIANFENG GAO, Microsoft Research Asia
and
CHIN-YEW LIN, Information Sciences Institute, University of Southern California

Categories and Subject Descriptors: 1.2.7 [**Artificial Intelligence**] Natural Language Processing – *Language models*

General Terms: Algorithms, Human Factors, Languages, Theory

Additional Key Words and Phrases: Statistical language modeling, n-gram models, discriminative training, source-channel models

1. INTRODUCTION

The goal of statistical language modeling (SLM) is to estimate the likelihood (or probability) of a word string. SLM is fundamental to many natural language applications like automatic speech recognition (ASR) [Jelinek 1990], statistical machine translation (SMT) [Brown et al. 1993], and Asian language text input [Gao et al. 2002a].

The research on SLM basically involves two main tasks: modeling and estimation. Modeling is to determine the structure of a statistical model; estimation is to determine the free parameters of the model using training data. SLM usually uses a parametric model with Maximum Likelihood Estimation (MLE) and various smoothing methods to tackle data sparseness problems. Different statistical models have been proposed in the past, but *n*-gram models (in particular, *bigram* and *trigram* models) still dominate SLM research.

SLM has recently been demonstrated as an effective framework for a few new applications, such as question answering [Berger 2001], text summarization, paraphrasing [Barzilay and Lee 2004], and information retrieval [Croft and Lafferty 2003]. However, these new applications come with new challenges. For example, in the SLM approaches to information retrieval, a language model has to be trained on a single document, an extremely small training set; while in ASR, a language model is typically trained on a million word corpus. The recent development of related techniques stimulates new modeling and estimation methods that are beyond the scope of the traditional approaches. Two representative examples of such techniques are statistical parsing and discriminative training.

With the ever-increasing popularity of SLM, we think that it is the right time to assemble a special issue reflecting recent advances in both its theory and applications. It

Authors' addresses: Jianfeng Gao, Microsoft Research Asia, 5F, Beijing Sigma Center. 100080, P. R. China. Email: jfgao@microsoft.com. Chin-Yew Lin, USC/Information Sciences Institute, 4767 Admiralty Way, Marina del Rey, CA 90292, USA. Email: cyl@isi.edu.

is our hope that the five articles in this special issue will enhance our understanding of the core problems in traditional SLM approaches and shed light on new principles, concepts, models, and applications for future work on SLM.

In Section 2, we briefly introduce applications using SLM; in Section 3, we give an overview of the theoretical background of SLM. We conclude this introduction with summaries of the five articles in this special issue. We do not provide a comprehensive survey, for which readers may refer to Chen and Goodman [1999]; Rosenfeld [2000]; and Goodman [2001].

2. APPLICATIONS

SLM initially attracted significant interests from the speech recognition community in the early 1980s [Jelinek 1990]. Statistics-based automatic speech recognition systems determine the word sequence W^* that best corresponds to an acoustic signal by solving the following equation:

$$W^* = \arg\max_{W \in \mathbf{GEN}(A)} P(W \mid A) = \arg\max_{W \in \mathbf{GEN}(A)} P(W)P(A \mid W), \quad (1)$$

where A denotes the input acoustic signal, and $\mathbf{GEN}(A)$ denotes the set of all possible word sequences that are acoustically similar to A. Equation (1) is called the source-channel models. We call the first term on the right-hand-side, the source model $P(W)$, the language model. In this framework, we view the desired word sequence W as generated by some source with probability $P(W)$ and transmitted through a noisy channel that transforms the intended W to the observation A with probability $P(A|W)$. The source-channel models have been widely used in many natural language processing (NLP) applications, and as a consequence, statistical language modeling also finds its way into many applications beyond the original ASR application. Depending on the application, language models are used to estimate the likelihood of a sequence of different types, such as parts-of-speech (POS) sequence in tagging [Church 1988]; source-language sentences in SMT [Brown et al. 1993]; correct word sequences in a spell-checker [Brill and Moore 2000]; and fluency of compressed English sentences in text summarization [Knight and Marcu 2002].

The use of SLM in the source-channel framework follows the principle of MLE: the source distribution is characterized by a language model (for example, an n-gram model), and the acceptability of an event (sequence) is measured by its likelihood of being generated from that distribution. Besides using a language model as a method of selecting high-quality source sequences, it can also be used solely as a ranking function. Competing events are ranked by their language model probabilities (or any transformation of the probabilities, provided that the transformation is order-preserving). The idea of using a language model as a ranking function has inspired NLP researchers to adopt the SLM approach in many new NLP tasks, and shown comparable or even better empirical results than conventional approaches. The application of SLM in information retrieval (IR) is a typical example, presented below.

The task of IR is to retrieve a ranked list of relevant documents D given a query Q. Unlike the classical probabilistic approach in which documents are ranked according to the probability of relevance (for example, the BIR model described in Jones et al. [1998]), in a SLM approach to IR, the retrieved documents are ranked by the posterior probability $P(D|Q)$, i.e., the probability that D is generated from the observed Q. By applying Bayes rule and dropping the constant denominator, we get $P(D|Q) \propto P(D)P(Q|D)$. We now have the source-channel models for IR. In practice, we usually

assume a uniform distribution of the prior probability $P(D)$, so the ranking function only takes $P(Q|D)$ into account. Since it is very difficult to estimate $P(Q|D)$ directly, we usually approximate $P(Q|D)$ by $P(Q|M_D)$, where M_D is the language model trained on D. In experiments, a language model is estimated for each document. Since the document is usually too small to train a reliable model, smoothing is one of the most critical issues. Zhai and Lafferty [2001] present an extensive empirical study of smoothing methods for SLM in IR applications. Also, owing to the sparse-data problem, most state-of-the-art SLM approaches to IR use unigram models and do not consider any dependency between words (e.g., Ponte and Croft [1998]; Zhai and Lafferty [2002]). Recently, there have been several attempts to capture word dependencies, with substantial improvements reported for large-scale experiments (e.g., Song and Croft [1999]; Miller et al. [1999]; Gao et al. [2004]). Hence the results demonstrate the possibility of introducing more advanced SLM techniques to IR.

Along with the development of new language modeling techniques in new applications, simple traditional language models such as the *n*-gram have also been applied in novel applications. For example, Barzilay and Lee [2004] used simple bigram language models trained on sentence clusters for different topics in a domain-specific document collection (for example, earthquakes, conflicts, and finance) to capture reoccurring topic sequences in a particular domain.

In the next section, we first give a brief overview of the theoretical background for SLM and then focus our discussion on two popular general models: the generative model and the discriminative model.

3. THEORETICAL BACKGROUND

The traditional approach to SLM, which uses an *n*-gram-based parametric model with MLE, is optimal in theory under two assumptions: (1) that the true distribution of data on which the parametric model is based is known; and (2) that there is sufficient training data. Unfortunately, these assumptions rarely hold in real-world scenarios.

In a traditional approach, a language model is typically a multinomial distribution, where an *n*-gram approximation is usually introduced to make the model tractable. For example, a trigram model predicts the current word, depending only on the two immediately preceding words. However, there are many cases in natural language where the independence assumption embedded in a trigram model does not hold. The distribution of words in a sentence is usually constrained by phrase structures that may relate two or more words over arbitrary distances. Due to the mismatch between the *approximate* distribution form (which is an *n*-gram model) and the *true* data distribution (which is unknown), the result of the MLE process is suboptimal. The deficiency has motivated a lot of research on SLM that is beyond the *n*-gram and MLE-based approaches. We outline below two research directions within the frameworks of the generative and discriminative models, respectively.

3.1 GENERATIVE MODELS

The research here is to produce a better (and usually more sophisticated) model to describe the generation process of text data more precisely by taking into account long-distance word dependencies. Such approaches can be classified along the scale of how much linguistic structure they use.

At one end of the spectrum, there are higher-order *n*-gram models and skipping models, which use no or very little linguistic information. However, these models are often much more complex than the traditional *n*-gram models and have too many free parameters to be estimated. As a result, little improvement has been reported, even with sophisticated smoothing techniques [Jelinek 1990; Goodman 2001].

At the other end of the spectrum, we have models that take syntactic structures, as well as lexical information, into account. The basic approach is to extend the conventional language model $P(W)$ to $P(W, T)$, where T is a parse tree of W. The extended model can then be used as a parser to select the most likely parse structure by $T^* = \text{argmax}_T P(W, T)$. Many recent studies [Chelba and Jelinek 2000; Charniak 2001; Roark 2001; Xu et al. 2002] adopt this approach. Similarly, dependency-based models [Lafferty et al. 1992; Collins 1996; Chelba et al. 1997; Gao and Suzuki 2003] use a dependency structure D of W instead of a parse tree T, where D is extracted from syntactic trees. Both models can be called *grammar-based models,* since they model the syntactic structure of a sentence. The models' parameters are estimated from syntactically annotated corpora such as the Penn Treebank. Most of these researches demonstrate an interesting and successful meld of SLM and statistical parsing techniques. Some significant improvement over *n*-gram models has been reported, although the proposed models are usually too computationally expensive to be deployed practically.

There are also models that fall between the two ends. For example, Gao et al. [2002b]; Isotani and Masunaga [1996] proposed linguistically-motivated skipping models where a *content* word is predicted, based on two previous *content* words that did not have to be adjacent (i.e., *function* words between content words are skipped). These models can be viewed as a simplified version of the grammar-based models, and are of more appealing practical value.

3.2 DISCRIMINATIVE MODELS

An alternative to the generative model is to explore the discriminative training approach to SLM, inspired by the observation that, in many applications, a language model is used as a ranking (or classification) function to discriminate between alternative solutions. This approach uses a general framework of linear models, derived from linear discriminant functions widely applied to pattern classification [Duda et al. 2001], and has recently been introduced into NLP tasks [Collins 2004]. Taking the ASR as an example, in the linear model framework, we have a set of $D + 1$ features $f_d(A, W)$, for $d = 0 \ldots D$. The features could be arbitrary functions that map (A, W) to real values. Using vector notation, we have $\mathbf{f}(A, W) \in \Re^{D+1}$. For each feature function, there is a model parameter λ_i. The best word string W^*, given the input acoustic signal A, can be determined by the decision rule as follows:

$$W^* = \arg\max_{W \in \mathbf{GEN}(A)} \boldsymbol{\lambda} \mathbf{f}(A, W) = \arg\max_{W \in \mathbf{GEN}(A)} \sum_{i=0}^{D} \lambda_i f_i(A, W) \tag{2}$$

Compared to the generative models, this approach has several appealing theoretical and practical properties. First, unlike the traditional approach, which depends heavily on the closeness of the estimate of *true* distributions, the discriminative training approach uses a much weaker assumption, i.e., that training and test data are generated from the same distribution, but that the form of the distribution (or the model structure) is unknown. Hence the above-mentioned problems of the traditional approaches, based on generative models, can be alleviated.

Second, the linear models provide a very flexible framework in which arbitrary linguistic knowledge sources can be incorporated by defining appropriate feature functions. For example, the source-channel models of Equation (1) can be viewed as a special case if we define both the source model and the channel model as two feature functions. In theory, by selecting these feature functions judiciously and letting D be sufficiently large, we can approximate any desired discriminant function. This framework

is conceptually similar to the un-normalized maximum-entropy language model described in Chen et al. [1998] and Rosenfeld [1994].

Finally, parameters are estimated using some discriminative training methods. A general approach is to create some criterion function and apply the gradient descent method. It is preferable to choose a criterion function that directly optimizes the performance measure of an application (e.g., to minimize the recognition word-error rate in ASR). Thus, discriminative training methods would potentially lead to better solutions, as shown in Roark et al. [2004].

In general, we prefer discriminative models to generative ones. Intuitively, if our goal is to discriminate between events A and B, it is enough to find the desired features that can differentiate them directly (as in discriminative models); while it is not necessary to first estimate the distributions of A and B, and then use the estimated distributions to construct the desired features (as in generative models). As pointed out by Vapnik [1998]: "when solving a given problem, solve it directly and try to avoid solving a more general problem as an intermediate step."

4. ARTICLES IN THIS ISSUE

The five articles in this special issue cover many aspects of SLM theory and applications. The first two address two fundamental problems of traditional SLM approaches, as we have discussed in Section 3: (1) how to obtain sufficient training data; and (2) how to construct a better model that can incorporate long-distance word dependencies.

Kim and **Khudanpur** propose two techniques to take advantage of in-domain text data in other languages. The authors assume that text data in a resource-rich language (source language) could be explored in developing models for another language (target language) without an explicit translation between the two languages. The two techniques are lexical triggers and latent semantic indexing, both have been applied to SLM and IR by other researchers, but are extended to cross-lingual context in this article. More interestingly, using the proposed techniques, a stochastic translation lexicon can be obtained on a document-aligned corpus rather than sentence-aligned corpus as in traditional MT approaches.

The language models presented by **Liares**, **Bendei**, and **Sanchez** fall into the category of grammar-based models, as described in Section 3.1. The proposed method follows a general paradigm: a combined model is constructed by linearly interpolating a general word-based trigram model and a grammar-based model (i.e., a category-based stochastic context-free grammar (SCFG)) that could capture long-distance dependencies between words. Although only a marginal improvement in the word-error rate is reported, the article presents an interesting method for learning a SCFG in general format and in Chomsky normal form in an (semi-)supervised manner. In particular, an initial SCFG is derived from the syntactic structures in the Penn Treebank corpus. Hence some well-known learning algorithms (i.e., the inside-outside method and its variants [Lari and Young 1990; Stolcke 1995]) can be performed on a (partially) bracketed corpus, which has proved to be highly effective [Pereira and Schabes 1992].

The next two articles present applications of SLM for tasks of spoken document retrieval and text summarization, respectively.

Chen, **Wang**, and **Lee** extend the SLM approach to IR, as described in Section 2, to Mandarin spoken-document retrieval. There are two major extensions: first, the language model is trained at both the word and syllable levels; second, the discriminative training method is introduced to improve the performance of document retrieval, which is formulated as a ranking (or classification) problem.

Nguyen, **Horiguchi**, **Shimazu**, and **Ho** present a novel application of SLM to text summarization. They first formulate text summarization as the sentence reduction

problem: that is, reducing long sentences to short ones while keeping the gist intact. Sentences are then reduced using template rules. The challenge is how to select the best combination of rule sequences from a large set of equally applicable rules that will lead to the best sentence reduction. This turns out to be a similar problem to that in ASR, where source-channel models can be applied. In this article, an *n*-gram model is used to estimate the likelihood of a rule sequence.

The last article in this issue, by **Fung**, **Ngai**, **Yang**, and **Chen**, presents a maximum-entropy Chinese parser. The article does not address the SLM problem itself, but instead reports the recent advances in a closely related technique, statistical parsing, on top of which more sophisticated SLM approaches can be developed. This article is also distinct from previous work on maximum-entropy English parsers [Ratnaparkhi 1998] in that some unique Chinese language problems (e.g., word segmentation) are investigated.

We hope that this special issue will trigger other interesting ideas for the development and use of SLM. We thank all the reviewers for their superb contributions. We also thank the TALIP editorial board, especially Kam-Fai Wong and Chang-Ning Huang, for helping us through all aspects of the selection process.

REFERENCES

BARZILAY, R. AND LEE, L. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *HLT-NAACL 2004*. 113-120.

BERGER, A. 2001. Statistical machine learning for information retrieval. Ph.D. dissertation. CMU. CMU-CS-01-110.

BRILL, E. AND MOORE, R. C. 2000. An improved error model for noisy channel spelling correction. In *ACL 2000*.

BROWN, P., PIETRA, S. D., PIETRA, V. D., AND MERCER, R. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics 19*, 2 (1993), 269-311.

CHARNIAK, E. 2001. Immediate-head parsing for language models. In *ACL/EACL 2001*. 124-131.

CHELBA, C. AND JELINEK, F. 2000. Structured language modeling. *Computer Speech and Language 14,* 4 (2000), 283-332.

CHELBA, C., ENGLE, D., JELINEK, F., JIMENEZ, V., KHUDANPUR, S., MANGU, L., PRINTZ, H., RISTAD, E. S., ROSENFELD, R., STOLCKE, A., AND WU, D. 1997. Structure and performance of a dependency language model. In *Processing of Eurospeech*, vol. 5. 2775-2778.

CHEN, S. F. AND GOODMAN, J. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language 13* (Oct. 1999), 359-394.

CHEN, S. F., SEYMORE, K., AND ROSENFELD, R. 1998. Topic adaptation for language modeling using unnormalized exponential models. In *ICASSP-98*.

CHURCH, K. W. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *ANLP 1988*. 136-143

COLLINS, M. J. 1996. A new statistical parser based on bigram lexical dependencies. In *ACL 34.* 184-191.

COLLINS, M. J. 2004. Parameter estimation for statistical parsing models: Theory and practice of distribution-free methods. In *New Developments in Parsing Technology*. H. Bunt et al. eds., Kluwer Academic, Amsterdam.

CROFT, W. B. AND LAFFERTY, J. (EDS.) 2003. *Language Modeling for Information Retrieval.* Kluwer Academic, Amsterdam.

DUDA, R. O., HART, P. E., AND STORK, D. G. 2001. *Pattern Classification*. Wiley, New York.

GAO, J., GOODMAN, J., LI, M., AND LEE, K. F. 2002a. Toward a unified approach to statistical language modeling for Chinese. *ACM Trans. on Asian Language Information Processing 1*,1 (March 2002), 3-33.

GAO, J., SUZUKI, H., AND WEN, Y. 2002b. Exploiting headword dependency and predictive clustering for language modeling. In *EMNLP 2002*. 248-256.

GAO, J., NIE, J. Y., WU, G., AND CAO, G. 2004. Dependence language model for information retrieval. In *SIGIR-2004* (Sheffield, UK, July 25-29, 2004),170-177.

GAO, J. AND SUZUKI, H. 2003. Unsupervised learning of dependency structure for language modeling. In *ACL 2003*. 521-528.

GOODMAN, J. T. 2001. A bit of progress in language modeling. *Computer Speech and Language* (Oct. 2001), 403-434.

ISOTANI, R. AND MATSUNAGA, S. 1994. A stochastic language model for speech recognition integrating local and global constraints. In *ICASSP-94*. 5-8.

JELINEK, F. 1990. Self-organized language modeling for speech recognition. In *Readings in Speech Recognition*, A. Waibel and K. F. Lee, eds. Morgan-Kaufmann, San Mateo, CA,1990, 450-506.

KNIGHT, K. AND MARCU, D. 2002. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence 139* (2002*),* 91-107.

LARI, K. AND YOUNG, S. J. 1990. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language 4* (1990), 35-56.

JONES, K. S., WALKER, S., AND ROBERTSON, S. 1998. A probabilistic model of information retrieval: Development and status. Tech. Rep. TR-446, Cambridge Univ. Computer Laboratory.

LAFFERTY, J., SLEATOR, D., AND TEMPERLEY, D. 1992. Grammatical trigrams: A probabilistic model of link grammar. In *Proceedings of the 1992 AAAI Fall Symposium on Probabilistic Approaches to Natural Language*.

MILLER, D. H., LEEK, T., AND SCHWARTZ, R. 1999. A hidden Markov model information retrieval system. In *SIGIR'99*.

PEREIRA, F. AND SCHABES, Y. 1992. Inside-outside reestimation from partially bracketed corpora. In *ACL 30*. 128-135.

PONTE, J. AND CROFT, W. B. 1998. A language modeling approach to information retrieval. In *SIGIR'98*. 275-281.

RATNAPARKHI, A. 1998. Maximum entropy models for natural language ambiguity resolution. Ph.D. dissertation, Univ. of Pennsylvania, Philadelphia, 1998.

ROARK, B. 2001. Probabilistic top-down parsing and language modeling. *Computational Linguistics 17*, 2 (2001), 1-28.

ROARK, B., SARACLAR, M., COLLINS, M. J., AND JOHNSON, M. 2004. Discriminative language modeling with conditional random fields and the perceptron algorithm. In *ACL 2004*.

ROSENFELD, R. 2000. Two decades of statistical language modeling: Where do we go from here? *Proc. IEEE 88* (Aug. 2000), 1270-1278.

ROSENFELD, R. 1994. Adaptive statistical language modeling: a maximum entropy approach. Ph.D. dissertation, Carnegie Mellon Univ., Pittsburgh, PA.

SONG, F. AND CROFT, W. B. 1999. A general language model for information retrieval. In *CIKM'99*, 316–321.

STOLCKE, A. 1995. An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics 21* (1995),165-202.

VAPNIK, V. N. 1998. *Statistical Learning Theory*. Wiley, New York.

XU, P., CHELBA, C., AND JELINEK, F. 2002. A study on richer syntactic dependencies for structured language modeling. In *ACL 2002*. 191-198.

ZHAI, C. AND LAFFERTY, J. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR2001*. 334-342.

ZHAI, C. AND LAFFERTY, J. 2002. Two-stage language models for information retrieval. In *SIGIR2002*. 49-56.

Received September 2004.