An Information-Retrieval Approach to Language Modeling: Applications to Social Data

Juan M. Huerta
IBM T. J.Watson Research Center
1101 Kitchawan Road
Yorktown Heights, NY 10598, USA
huerta@us.ibm.com

Abstract

In this paper we propose the IR-LM (Information Retrieval Language Model) which is an approach to carrying out language modeling based on large volumes of constantly changing data as is the case of social media data. Our approach addresses specific characteristics of social data: large volume of constantly generated content as well as the need to frequently integrating and removing data from the model.

1 Introduction

We describe the Information Retrieval Language Model (IR-LM) which is a novel approach to language modeling motivated by domains with constantly changing large volumes of linguistic data. Our approach is based on information retrieval methods and constitutes a departure from the traditional statistical n-gram language modeling (SLM) approach. We believe the IR-LM is more adequate than SLM when: (a) language models need to be updated constantly, (b) very large volumes of data are constantly being generated and (c) it is possible and likely that the sentence we are trying to score has been observed in the data (albeit with small possible variations). These three characteristics are inherent of social domains such as blogging and micro-blogging.

2 N-gram SLM and IR-LM

Statistical language models are widely used in main computational linguistics tasks to compute the probability of a string of words: $p(w_1...w_i)$
To facilitate its computation, this probability is expressed as:

$$p(w_1...w_i) = P(w_1) \times P(w_2 \mid w_1) \times ... \times P(w_i \mid w_1...w_{i-1})$$

Assuming that only the most immediate word history affects the probability of any given word, and focusing on a trigram language model:

$$P(w_i \mid w_1...w_{i-1}) \approx P(w_i \mid w_{i-2}w_{i-1})$$

This leads to:

$$P(w_1...w_i) \approx \prod_{k=1..i} p(w_k \mid w_{k-1}w_{k-2})$$

Language models are typically applied in ASR, MT and other tasks in which multiple hypotheses need to be rescored according to their likelihood (i.e., ranked). In a smoothed backoff SLM (e.g., Goodman (2001)), all the n-grams up to order n are computed and smoothed and backoff probabilities are calculated. If new data is introduced or removed from the corpus, the whole model, the counts and weights would need to be recalculated. Levenberg and Osborne (2009) proposed an approach for incorporating new data as it is seen in the stream. Language models have been used to support IR as a method to extend queries (Lavrenko et al. 2001); in this paper we focus on using IR to carry out language modeling.

2.1 The IR Language Model

The IR-LM approach consists of two steps: the first is the identification of a set of matches from a corpus given a query sentence, and second is the estimation of a likelihood-like value for the query.

In the first step, given a corpus *C* and a query sentence *S,* we identify the k-closest matching sentences in the corpus through an information retrieval approach. We propose the use of a modified String Edit Distance as score in the IR process. To efficiently carry out the search of the closest sentences in the corpus we propose the use of an inverted index with word position

Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, pages 7–8,
Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

information and a stack based search approach described in Huerta (2010). A modification of the SED allows queries to match portions of long sentences (considering local insertion deletions and substitutions) without penalizing for missing the non-local portion of the matching sentence.

In the second step, in general, we would like to compute a likelihood-like value of S through a function of the distances (or alternatively, similarity scores) of the query S to the top k-hypotheses. However, for now we will focus on the more particular problem of ranking multiple sentences in order of matching scores, which, while not directly producing likelihood estimates it will allow us to implement n-best rescoring. Specifically, our ranking is based on the level of matching between each sentence to be ranked and its best matching hypothesis in the corpus. In this case, integrating and removing data from the model simply involve adding to or pruning the index which generally are simpler than n-gram re-estimation.

There is an important fundamental difference between the classic n-gram SLM approach and our approach. The n-gram approach says that a sentence S_1 is more likely than another sentence S_2 given a model if the n-grams that constitute S_1 have been observed more times than the n-grams of S_2. Our approach, on the other hand, says that a sentence S_1 is more likely than S_2 if the closest match to S_1 in C resembles S_1 better than the closes match of S_2 resembles S_2 regardless of how many times these sentences have been observed.

3 Experiments

We carried out experiments using the blog corpus provided by Spinn3r (Burton et al (2009)). It consists of 44 million blog posts that originated during August and September 2008 from which we selected, cleaned, normalized and segmented 2 million English language blogs. We reserved the segments originating from blogs dated September 30 for testing.

We took 1000 segments from the test subset and for each of these segments we built a 16-hypothesis cohort (by creating 16 overlapping sub-segments of the constant length from the segment).

We built a 5-gram SLM using a 20k word dictionary and Knesser-Ney smoothing using the SRILM toolkit (Stolcke (2002)). We then ranked each of the 1000 test cohorts using each of the model's n-gram levels (unigram, bigram, etc.). Our goal is to determine to what extent our approach correlates with an n-gram SLM-based rescoring.

For testing purposes we re-ranked each of the test cohorts using the IR-LM approach. We then compared the rankings produced by n-grams and by IR-LM for every n-gram order and several IR configurations. For this, we computed the Spearman rank correlation coefficient (SRCC). SRCC averages for each configuration are shown in table 1. Row 1 shows the SRCC for the best overall IR configuration and row 2 shows the SRCC for the IR configuration producing the best results for each particular n-gram model. We can see that albeit simple, IR-LM can produce results consistent with a language model based on fundamentally different assumptions.

	n=1	n=2	n=3	n=4	n=5
overall	0.53	0.42	0.40	0.40	0.38
individual	0.68	0.47	0.40	0.40	0.39

Table 1. Spearman rank correlation coefficient for several n-gram IR configurations

4 Conclusion

The IR-LM can be beneficial when the language model needs to be updated with added and removed data. This is particularly important in social data where new content is constantly generated. Our approach also introduces a different interpretation of the concept of likelihood of a sentence: instead of assuming the frequentist assumption underlying n-gram models, it is based on sentence feasibility based on the closest segment similarity. Future work will look into: integrating information from the top k-matches, likelihood regression, as well as leveraging other approaches to information retrieval.

References

Burton K., Java A., and Soboroff I. (2009) The ICWSM 2009 Spinn3r Dataset. *Proc. ICWSM 2009*

Goodman J. (2001) A Bit of Progress in Language Modeling, *MS Res. Tech. Rpt. MSR-TR-2001-72.*

Huerta J. (2010) A Stack Decoder Approach to Approximate String Matching, *Proc. of SIGIR 2010*

Lavrenko V. and Croft W. B. (2001) Relevance based language models. *Proc. of SIGIR 2001*

Levenberg A. and Osborne M. (2009), Stream-based Randomised Lang. Models for SMT, *EMNLP 2009*

Stolcke A. (2002)s SRILM -- An Extensible Language Modeling Toolkit. *Proc. ICSLP 2002*