USING CROSS-LANGUAGE CUES FOR STORY-SPECIFIC LANGUAGE MODELING

Sanjeev Khudanpur and Woosung Kim

Center for Language and Speech Processing
The Johns Hopkins University, Baltimore, MD 21218
{sanjeev,woosung}@clsp.jhu.edu

ABSTRACT

We propose methods to exploit contemporary news articles in a resource rich language, together with cross-language information retrieval and machine translation, to sharpen language models for a news story in a language with fewer linguistic resources. We report experimental results on story-specific Chinese language models that use cues from a parallel corpus of English news stories. We demonstrate that even with fairly crude cross-language information retrieval, level-1 machine translation and simple linear interpolation, a significant (18%) reduction in perplexity may be obtained over a Chinese trigram model. We also demonstrate that this method of sharpening the Chinese language model is complementary to other techniques like topic dependent modeling, and the two in combination result in an even greater reduction in perplexity (28%).

1. INTRODUCTION

The last decade has seen a dramatic improvement in the capability and performance of automatic systems for processing speech and natural language. This progress may be attributed largely to advances in statistical modeling techniques and procedures for automatic learning from large speech and text corpora. The construction of increasingly accurate and complex stochastic models, in particular, is crucially dependent on the availability of large corpora of transcribed speech and on-line text specific to the language and the application domain. Much of these advances, therefore, have been in languages such as English, French, German and Japanese, and in domains such as air travel information and broadcast news transcription, for which such linguistic resources have been created at considerable cost.

Construction of accurate stochastic models for processing resource deficient languages has recently started receiving attention. Methods have been proposed to bootstrap acoustic models for automatic speech recognition (ASR) in resource deficient languages by reusing acoustic models from resource-rich languages — the notion of a universal phone-set (cf e.g. [1]) has been used to jointly train acoustic models in multiple languages, or phonetic models in the target language have been “synthesized” by matching well-trained models from resource-rich languages to a limited amount of transcribed speech in the target language [2]. Morphological analyzers, noun-phrase chunkers, part-of-speech taggers etc., have been developed (cf e.g. [3]) for resource deficient languages by exploiting translated texts — statistical models are used to align words in a sentence in the target language with words in, say, the English translation of the sentence; the English side is automatically annotated for the necessary categories (POS tags, NP brackets), and the annotation is projected to the target language via the alignment, producing a “labeled” corpus in the resource deficient language, from which necessary statistical models are then estimated.

In this paper, we propose an approach for estimating a language model (LM) for ASR in resource deficient languages. When an ASR system needs to be engineered for a specific domain in a new language (e.g. Arabic Broadcast News), a modest amount of domain specific LM training text is usually made available, from which a word-list and a small N-gram model may be derived. Additional target language text from an unrelated domain (e.g. Arabic web pages) may sometimes be available, and its use to improve performance in the target language and domain has been investigated elsewhere [4]. Domain specific text in other languages (e.g. English Broadcast News) is also often available and in this paper we investigate methods to use such cross-language in-domain data to improve the LM.

The topic detection and tracking (TDT) task [5] is a concrete example of a large publicly funded technology demonstration program which motivates the research described in this paper. The TDT corpus contains news broadcasts from 4 audio sources and 2 text sources in English as well as 1 audio source and 2 text sources in Mandarin. The broadcasts have been collected concurrently over a 9 month period in

1A limited amount of linguistic resources can almost always be produced with moderate effort; resource deficiency here implies the lack of 100s of millions of words of text, 100s of thousands of manually parsed sentences, 10s of hours of orthographically transcribed speech, etc.

This research was partially supported by the National Science Foundation (via Grant No IIS-9982329) and the Office of Naval Research (via a MURI Grant).
1998. The goal of the TDT program is to demonstrate a system which can automatically track the reporting of a specified event or set of events across all the news sources, to detect new events as soon as they are reported in any one of the sources, etc. The audio sources are transcribed by language-specific ASR systems and the rest of the processing does not explicitly distinguish between speech and text sources. It has been noted in TDT literature that ASR errors, particularly those of named entities and infrequently occurring “content words,” degrade event detection and tracking performance [6].

It has been demonstrated that even with mediocre ASR, existing cross-language information retrieval (CLIR) techniques can be effectively employed to identify concurrent English news stories which are on the same topic as an audio story in a target, resource deficient language. In this paper we propose methods to exploit such contemporary English stories in conjunction with machine-translation to sharpen the language model for each individual audio news story in the target language, an exercise we call story-specific language modeling, as illustrated in Figure 1. A second-pass transcription of the audio with such sharper models may be employed to improve the ASR accuracy up from being adequate for CLIR to being usable for summarization or for information extraction.

In the following, we present details of this method, and report on our initial experiments in story-specific Chinese LMs using cues from contemporary English news stories. Section 2 describes specifics of our bilingual corpus of translated Chinese-English stories, and the experimental setup we have used. Section 3 presents our main experimental results and some analysis. We demonstrate a significant (18%) reduction in perplexity on the Chinese stories over a Chinese trigram LM, using only unigram cues from the English stories and an elementary interpolation method. Since the contribution of the contemporary English story is likely to be due to a sharper topic-focus, we compare our gains with those obtained by more traditional topic language modeling techniques in Section 3.1. Section 4 concludes with a discussion of our plans for further development and evaluation of story-specific language models.

2. STORY SPECIFIC LANGUAGE MODELING

We use the Hong Kong News parallel text corpus for all experiments reported in this initial investigation. The corpus contains 18,147 aligned Chinese-English article pairs, dating from July 1997 to April 2000, released by the Information Services Department of Hong Kong Special Administrative Region of the People’s Republic of China; through the Linguistic Data Consortium [7]. After removing a few articles containing nonstandard Chinese characters, we divide the corpus, by random selection, into 16,010 article-pairs for training, 750 pairs for cross-validation and other development, and 682 pairs for evaluation. Only perplexity and out-of-vocabulary (OOV) rate measurements are performed on the evaluation portion of the corpus; no parameters are tuned on it, nor any iterative diagnostics performed.

All the Chinese articles, training, development and evaluation sets included, have been automatically segmented into words (cf. [8]). This results in a 4.2M-word Chinese LM training set (CH-train), a 255K-word Chinese development set (CH-dev) and a 177K-word Chinese evaluation set (CH-eval). A vocabulary of 41K-words covering CH-train and CH-dev is used for language modeling and it results in a 0.4% OOV rate on CH-eval. By comparison, the English counterpart of our Chinese LM corpus contains 4.3M-words in the training set (EN-train), 263K-words in the development set (EN-dev) and 182K-words in the evaluation set (EN-eval); the English vocabulary size is 39K-words and the OOV rate on EN-eval is 0.4% — in harmony with the figures for the Chinese portion due to the fact that the article-pairs are indeed translations of each other.

2.1. Baseline Chinese Language Model Estimation

We estimate a standard trigram LM, using Good-Turing discounting and Katz back-off, from CH-train. Its perplexity on CH-dev and CH-eval is reported in Section 3. We considered basing all our Chinese LMs on character N-grams instead of words, but went with with a word-based LM primarily because we believe that the cross-language cues will be directly beneficial to a word based model.

Chinese LM discussions, particularly for ASR, frequently report character perplexity (instead of word perplexity) and character error rates, mainly to facilitate comparison across approaches that use different word-segmentations. We too report character perplexity: while calculating the average perplexity of a set of sentences, we simply divide the log-probability of a sentence by the number of characters in the sentence rather than the number of segmented words. All our models, however, assign probability to entire words.
2.2. The Cross-Language Language Model

Let \( dC_1, \ldots, dC_N \) denote the Chinese news articles in the test set (CH-dev or CH-eval) and let \( dE_1, \ldots, dE_N \) denote their corresponding aligned English news articles. Assume for the time being that this correspondence is known; we will demonstrate shortly that this is not difficult to discover automatically with reasonable accuracy. Assume further that we have at our disposal a “level-1” translation system — a probabilistic model of the form \( P(c|e) \) — which provides the Chinese translation \( c \) of each English word \( e \) from our Chinese and English vocabularies \( C \) and \( E \) respectively; we will also shortly demonstrate a way of obtaining such a model from the given training data. Given these two assumptions, it seems plausible that

\[
P(c|dE_i) = \sum_{e \in E} P(c|e) \hat{P}(e|dE_i), \quad \forall c \in C,
\]

would be a good unigram LM for the \( i \)-th Chinese article \( dC_i \), given the relative frequency \( P(\cdot|dE_i) \) of words in the corresponding English document.

We of course do not expect this unigram LM to outperform our baseline trigram LM. We therefore investigate the linear interpolation

\[
P(c|dE_i) = \lambda P(c|dC_i) + (1 - \lambda) P(c|dE_i|dC_i) \quad (2)
\]

of the cross-language unigram model with the baseline trigram model. The development set CH-dev is used to estimate the interpolation weight \( \lambda \).

2.3. Estimating Translation Models \( P(c|e) \) and \( P(e|c) \)

The Hong Kong News corpus was automatically aligned at the sentence level (cf. [8]) and we use GIZA++, a statistical machine translation training tool [9], to train an IBM-Model-3 translation system based on the 16K article pairs from our training set. We extract the translation tables \( P(c|e) \) and \( P(e|c) \) from GIZA++ and use them as level-1 translation models. We note that since we apply these translation models to entire English (or Chinese) articles, the requirement of a sentence aligned corpus is, in theory, not essential. Indeed, it is our claim that since \( P(c|e) \) and \( P(e|c) \) will be used for language modeling, it is not even essential that \( dC_i \) and \( dE_i \) be exact translations of each other; their being on the same topic or story ought to suffice.

3. EXPERIMENTAL RESULTS

We first assume that the alignment of a Chinese test document \( dC_i \) with its English counterpart \( dE_i \) is given, and compute the language model of equation (2) above. We report the average perplexity\(^2 \) for CH-dev and CH-eval in the second row of Table 1; both word- and character-perplexity are reported together for completeness.

Next, we relax the assumption that the story-alignment is given. For each Chinese article \( dC_i \), we use the reverse translation model \( P(e|c) \) described above to create an English bag-of-words representation as used in standard information retrieval (IR), and use it to find the English document with the highest cosine similarity — this document then plays the role of \( dE_i \). The results in rows 2 and 3 of Table 1 indicate that with reasonable CLIR performance, the cross language LM is robust to story alignment errors.

As an aside, the correct English document is retrieved from EN-dev for 92% of the articles in CH-dev, and from EN-eval for 89% of the articles in CH-eval. The EN-dev and EN-eval sets are small in size relative to IR document collections, but if one were looking at English newswire feed on a given day for an article to match a Chinese story, it ought to be feasible to narrow the search down to a few hundred candidate articles.

3.1. Comparison with Topic-Dependent LMs

It stands to reason that while the crude nature of our translation model \( P(c|e) \) and the cosine-based CLIR model used here are unlikely to result in extraction of any detailed knowledge from \( dE_i \) about the word-sequences in a target Chinese story \( dC_i \), it is probably through some mechanism of “topic focussedness” that the story specific LMs obtain a reduced perplexity. This intuition motivates us to construct topic-dependent LMs of the type discussed, e.g., in [10] and contrast their performance with the models of rows 2-3 in Table 1. We proceed as follows.

The 16K articles in CH-train are each represented by a bag-of-words vector using standard IR methods. These 16K vectors are then clustered into 100 classes using a standard K-means clustering algorithm. Random initialization is used to seed the algorithm, and cosine similarity is used as the “metric” for clustering. Five iterations of the K-means algorithm are performed, and the resulting 100 clusters are deemed to represent different “topics.” A bag-of-words vector (centroid) created from all the articles in a cluster is used to represent each topic. Topic specific unigram and trigram LMs, denoted \( P_1(t) \) and \( P_3(t|c_k|c_{k-1}, c_{k-2}) \) respectively, are also computed for each topic \( t \) exclusively from the articles in its cluster.

\(^2\text{Note that in results reported here, the entire article } dC_i \text{ is used to determine } dE_i \text{, which in turn conditions the probability assigned to words in } dC_i \text{. Strictly speaking, this is inappropriate conditioning of the probabilistic model. However, the theoretically correct version of conditioning the LM for each word } c_k \text{ only on } c_1, \ldots, c_{k-1}, \text{ is known from many other cases to produce nearly identical results — due to the robust determination of } dE_i — \text{ so we proceed with this somewhat “polluted” investigation.}
We recognize that while we have obtained substantial reductions in perplexity, it remains to be seen if this translated into improved performance in specific applications — ASR or machine translation. We expect to apply this model to Chinese ASR in the near future.

Finally, we also note that the Chinese-English corpus on which we have conducted these experiments provides certain advantages which may not be available in cases such as the TDT task, where the articles only discuss identical events, but aren’t translations of each other. We plan to conduct comparable experiments on the TDT task. We are optimistic that the minimal assumption made by our model, that the English story merely provides a sharper unigram distribution of Chinese words in the aligned story, will continue to hold even for aligned (but not translation equivalent) TDT story-pairs, and we expect comparable results.

5. REFERENCES


