23年 1月 14日

Electronic and Information Engineering Department	ID	065401
Name	WELLY NAPTALI	

Advisor	Seiichi Nakagawa Tomoyoshi Akiba

Title	Study on n-gram language models for topic and out-of-vocabulary words

(800 words)

Language models (LMs) are an important field of study in automatic speech recognition (ASR) systems. LM helps acoustic models find the corresponding word sequence of a given speech signal. Without it, ASR systems would not understand the language and it would be hard to find the correct word sequence. A data sparseness problem for modeling a language often occurs in LMs. The problem is caused by the insufficiency of training data, which in turn, makes the infrequent words have unreliable probability.

In this research, we investigated a class LM based on a latent semantic analysis (LSA). A word-document matrix is commonly used to represent a collection of text (corpus) in LSA framework. This matrix tells how many times a word occurs in a certain document. In other words, this matrix ignores the word order in the sentence. We propose several word co-occurrence matrices that keep the word order. By applying LSA to these matrices, words in the vocabulary are projected to a continues vector space according to their position in the sentences. To support these matrices, we define a context dependent class (CDC) based n-gram. Unlike traditional class-based n-gram LM, CDC LM distinguishes classes according to their context in the sentences. Experiments on Wall Street Journal (WSJ) corpus show that the word co-occurrence matrix works 3.62%~12.72% better in terms of perplexity than word-document matrix. Furthermore, the CDC improves the performance and achieves better perplexity than the traditional class-based n-gram LM based on LSA. When the model is linearly interpolated with the word-based 3-gram, it gives improvements about 2.01% for 3-gram model and 9.47% for 4-gram model on relative perplexity against a standard word-based 3-gram LM.

During the past few years, researchers have tried to incorporate long-range dependencies into statistical word-based n-gram LMs. One of these long-range dependencies is topic. Unlike words, topic is unobservable. Thus, it is required to find the meanings behind the words to get into the topic. As the second part of this research, we proposed a new approach for a topic-dependent LM called topic dependent class (TDC) based n-gram, where the topic is decided in an unsupervised manner. LSA is employed to reveal hidden (latent) relations among nouns in the context words. To decide the topic of an event, a fixed size word history sequence (window) is observed, and voting is then carried out based on noun class occurrences weighted by a confidence measure. Experiments were conducted on an English corpus and a Japanese corpus: WSJ corpus and Mainichi Shimbun (Japanese newspaper) corpus. The results show that our proposed method gives better perplexity than the comparative baselines; including a word-based/class-based n-gram LM, their interpolated LM, a cache-based LM, a topic mixture LM based on n-gram, and a topic mixture LM based on Latent Dirichlet Allocation (LDA). The TDC LM achieved a relative perplexity improvement over the word-based 3-gram of 14.0% and 15.2% for the WSJ and Mainichi Shimbun

corpora, respectively. The n-best list rescoring was conducted to validate its application in ASR systems.

In the third part of the research, we made some extensions to the TDC LM. The contribution is threefold. First, the TDC is improved further by performing soft-clustering and/or soft-voting techniques on the training and/or test phases, which solve a data shrinking problem and make TDC independent from the word-based n-gram LM. Second, we incorporate a cache-based LM through unigram scaling to obtain further improvements, since TDC and cache-based LM captures different property of the language. Finally, we provide the evaluation in terms of WER and analysis on an ASR rescoring task. Experiments on WSJ and Mainichi Shimbun show that TDC LM improves both perplexity and word-error-rate (WER). The perplexity reduction is up to 25.1% relative on English corpus and 25.7% relative on Japanese corpus. Furthermore, the best reduction on WER is 15.2% relative on English ASR and 24.3% relative on Japanese ASR compared to the baseline

In the last part of this research, we investigated an LM about out-of-vocabulary (OOV) words. OOV in a language model always occurs in real life application of a speech recognition system due to insufficient corpus or vocabulary limitation. Though they are infrequent, they are very important for real applications like Information Extraction, because most of them are proper nouns and new topic words. Inflected languages are the most suffered by an OOV problem. The common way to handle the problem is by changing the recognition basic unit from word-based to subword-based. In this research, we proposed a framework to estimate the probability of OOV word with the help of data taken from World Wide Web (WWW). This OOV word is used as a query to a search engine to get several web pages. From these web pages, we use a defined similarity measure to find the relations of this OOV word to other in-vocabulary (IV) words. Then a probability is assigned using a class-based language model. Preliminary experiments were conducted on WSJ corpus and showed the validness.