Language Model Adaptation

Special Topics in Spoken Language Processing

Guan-Yu Menphis Chen

Department of Computer Science & Information Engineering,
National Taiwan Normal University, Taipei, Taiwan
696470203@ntnu.edu.tw

Main Reference：
J. R. Bellegarda, “Statistical language model adaptation: review and perspectives”, 2003.

Outline

- Introduction
- Why Adaptation?
- Adaptation Framework
- Adaptation Techniques
  - Model Interpolation
  - Constraint Specification
  - Topic Information
  - Semantic Knowledge
  - Syntactic Infrastructure
  - Multiple Sources
- Conclusion
- Reference

Introduction

- Language modeling plays a pivotal role in automatic speech recognition.
- Language model used to constrain the acoustic analysis, guide the search through multiple hypotheses, and contribute to the determination of the final transcription.
- In the search, the successful capture of this information is critical to help determine the most likely sequence of words spoken, because it quantifies which word sequences are acceptable in a given language for a given task, and which are not.

Why Adaptation?

- Natural language is highly variable in several aspects.
  1. Language evolves as does the world it seeks to describe.
     - The effective underlying vocabulary changes dynamically with time on a constant basis.
  2. Different domains tend to involve relatively disjoint concepts with markedly different word sequence statistics.
     - For example, “interest rate” to a banking application is different to gaming platforms.
  3. People naturally adjust their use of the language based on the task at hand.
     - Compare the typical syntax employed in formal technical papers to the one in casual e-mails.
  4. People’s style of discourse may independently vary due to a variety of factors such as socio-economic status, emotional state, etc.

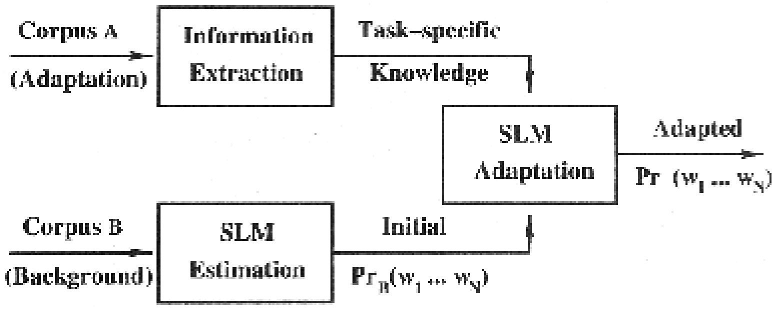
Why Adaptation?

- As a result of this inherent variability, the lexical, syntactic, or semantic characteristics of the discourse in the training and recognition tasks are quite likely to differ.

- Linguistic mismatch is known to affect cross-task recognition accuracy much more than acoustic mismatch. For instance, results of a cross-task experiment using Broadcast News models to recognize TI-digits show that only about 8% of the word error rate increase was due to the acoustic modeling mismatch, while 92% was imputable to the language model mismatch.

Adaptation Framework

- The general SLM (statistical language model) adaptation framework is depicted below. Two text corpora are considered: a (small) adaptation corpus A , and a (large) background corpus B .



- Given a sequence of N words w_q $(1 \le q \le N)$ consistent with the corpus A , the goal is to compute a suitably robust estimate of the language model probability：

$$Pr(w_1,\dots,w_N) = \prod_{q=1}^{N} Pr(w_q \mid h_q)$$

where h_q represents the history available at time q .

- For an n-gram model, the Markovian assumption implies $h = w_{q-n+1},\dots, w_{q-1}$.

Adaptation Framework

- The estimation of $Pr(w_1,\ldots,w_N)$ leverages two distinct knowledge sources：
  1. The well-trained, but possibly mismatched, background SLM.
  2. The adaptation data.

- The general idea is to dynamically modify the background SLM estimate on the basis of what can be extracted from A .

- In some cases, the corpus A may already be available. For instance, in cross-domain adaptation, a small amount of domain-specific text may have been collected for some other purpose, but can readily serve as adaptation data.

- On the other hand, if the corpus A is not available a priori, or deemed too small, appropriate text needs to be gathered. One approach that turns out to be quite efficient is to use multiple sentence hypotheses from an N-best list as adaptation material. Every sentence now contributes to the corpus according to its weight.

Adaptation Technique - Model Interpolation

- In interpolation-based approaches, the corpus is used to drive a task-specific (dynamic) SLM, which is then combined with the background (statistic) SLM.

- Model interpolation can be divided into three kinds：
  1. Model Merging
  2. Dynamic Cache Models
  3. MAP Adaptation

Adaptation Technique - Model Merging

- Because of the extremely limited amount of data involved, the dynamic SLM tends to be poorly trained, most of the time resulting in a rather inaccurate estimate.
- But for certain idiosyncratic word sequences, particularly frequent in the current task, may the dynamic model outperform the initial estimate SLM.
- The simplest way to do so is via linear interpolation.

$$Pr(w_q \mid h_q) = (1-\lambda) \cdot Pr_A(w_q \mid h_q) + \lambda \cdot Pr_B(w_q \mid h_q)$$

$$where\ 0 \le \lambda \le 1$$

- Alternatively, it is possible to back-off from the dynamic estimate to the static one depending on the associated frequency count.

$$Pr(w_q \mid h_q) = \begin{cases} Pr_A(w_q \mid h_q) & if\ C_A(h_q, w_q) \ge T \\ \beta\, Pr_B(w_q \mid h_q) & otherwise \end{cases}$$

Where T is an empirical threshold, and β is calculated to ensure that $Pr(w_q \mid h_q)$ is a true probability.

Adaptation Technique - Dynamic Cache Models

- A special case of linear interpolation, widely used for within-domain adaptation, deserves special mention：dynamic cache memory modeling.

- The idea underlying the model was that a language model that exploited short-term shifts in word-use frequencies might perform significantly.

- In an effort to propagate the power of the method to higher order cases, the cache memory concept has been extensively applied in conjunction with the class model of the form：

$$Pr\left(w_q \mid h_q\right) = \sum_{\{c_q\}} Pr\left(w_q \mid c_q\right) Pr\left(c_q \mid h_q\right)$$

  where $\{c_q\}$ is a set of possible classes for word w_q , given the current history h_q .

Adaptation Technique - Dynamic Cache Models

- The language model thus comprises a class n-gram component $Pr(c_q \mid h_q)$ and a class assignment component $Pr(w_q \mid c_q)$.
- The class n-gram component is taken from the background SLM, i.e., $Pr(c_q \mid h_q) = Pr_B(c_q \mid h_q)$.
- The class assignment component is subject to dynamic cache adaptation, resulting in $Pr(w_q \mid c_q) = (1 - \lambda) \cdot Pr_A(w_q \mid c_q) + \lambda \cdot Pr_B(w_q \mid c_q)$.

Adaptation Technique - MAP Adaptation

- Recently, it is argued that combination should be done at the frequency count level rather than the model level.

- In the approach, the MAP-optimal model M^* is computed as

  $$M^* = argmax_M Pr(A \mid M)Pr(M)$$

  where $Pr(M)$ is a prior distribution over all models.

- The framework leads to a solution of the form：

  $$Pr\left(w_q \mid h_q\right) = \frac{\varepsilon C_A\left(h_q w_q\right) + C_B\left(h_q w_q\right)}{\varepsilon C_A\left(h_q\right) + C_B\left(h_q\right)}$$

  where ε is a constant factor which is estimated empirically.

Adaptation Technique - Constraint Specification

- In approaches based on constraint specification, the corpus is used to extract features that the adapted SLM is constrained to satisfy.

- This is arguably more powerful than model interpolation, since in this framework a different weight could presumably be assigned separately for each feature.

- We will discuss：
  1. Exponential Model
  2. MDI Adaptation
  3. Unigram Constraints

Adaptation Technique - Exponential Model

- Constraint-based methods have been associated with exponential models trained using the maximum entropy (ME) criterion.

- Rather than deriving the conditional probability $Pr(w|h)$ directly, consider the associated event of the joint probability distribution.

- Assume that this joint distribution is constrained by K linearly independent constraints, written as

$$\sum_{\{(h,w)\}} I_k(h,w)Pr(h,w) = \alpha(\hat{h}_k \hat{w}_k),\, 1 \le k \le K$$

where I_k is the indicator function of an appropriate subset of the sample space, and $\alpha(\hat{h}_k \hat{w}_k)$ denotes the relevant empirical marginal probability.

- It can be show that the joint distribution $Pr(w|h)$ has the parametric form：

$$Pr(w|h) = \frac{1}{Z(h,w)} \prod_{k=1}^{K} exp\{\lambda_k I_k(h,w)\}$$

Adaptation Technique - MDI Adaptation

- In MDI (minimum discrimination information estimation), the features extracted from corpus A are considered as important properties.

- The solution has to be close to the joint background distribution $Pr_B(h,w)$, taken as prior distribution. This is achieved by minimizing the Kullback-Leibler distance from the joint background distribution：

  $$\min_{Q(h,w)} \sum_{\{(h,w)\}} Q(h,w) log \frac{Q(h,w)}{Pr_B(h,w)}$$

  while simultaneously satisfying the linear constraints：

  $$\sum_{\{(h,w)\}} I_k(h,w) Q(h,w) = \alpha_A(\hat{h}_k, \hat{w}_k), \ 1 \le k \le K$$

  where α_A emphasizes the fact that the relevant empirical probabilities obtained from the adaptation corpus A .

- It can be show that the solution has the form：

  $$\Pr(h,w) = \frac{\Pr_B(h,w)}{Z(h,w)} \prod_{k=1}^{K} \exp\{\lambda_k I_k(h,w)\}$$

Adaptation Technique - Unigram Constraints

- An important special case is MDI adaptation with unigram constraints. Given the typically small amount of adaptation data, it is often the case that only unigram features can be reliably estimated on the adaptation corpus A .

- In the case, the constraints become：

$$\sum_{\{(h,w)\}} I_k(h,w)Q(h,w) = \alpha_A(\hat{w}_k),\ 1 \le k \le K$$

where $\alpha_A(\hat{w}_k)$ now represents the empirical unigram probability obtained from A for the feature $\hat{w}_k$.

- And in fact it can be shown that the resulting solution reduces to the closed form：

$$\Pr(h,w) = \Pr_B(h,w)\frac{\alpha_A(w)}{\Pr_B(w)}$$

Adaptation Technique - Topic Information

- The approaches exploiting the topic information about the underlying subject matter from corpus A . The information is used to improve the background model based on semantic classification.

- We will discuss：
  1. Mixture Model
  2. Explicit Topic Model

Adaptation Technique - Mixture Model

- The simplest approach is based on a generalization of linear interpolation to include several pre-defined domain.

- Assume the background n-gram model is composed of a collection of K sub-models, each trained on a separate topic. Mixture SLMs linearly interpolate these K n-grams in such a way that the resulting mixture best matches the adaptation data A .

- The probability is obtained as：
$$Pr\left(w_q \mid h_q\right)=\sum_{k=1}^{K} \lambda_{A,k}\, Pr_{B,k}\left(w_q \mid h_q\right)$$
where $Pr_{B,k}$ refers to the $k\text{-}th$ pre-defined topic sub-model, and the notation $\lambda_{A,k}$ for the interpolation coefficients reflects the fact that they are estimated on A .

- It turns out that, in actual usage, the mixture SLM is less practical than a single SLM, in part because it complicates smoothing.

Adaptation Technique - Explicit Topic Model

- Mixture modeling includes topic information indirectly. Another solution is to express the topic contribution more directly.

- Consider the language model probability $Pr\left(w_q \mid h_q\right)=\sum_{k=1}^{K} Pr\left(w_q \mid t_k\right) Pr\left(t_k \mid h_q\right)$ where t_k is one of the K topics above.

- There is no assumption that each history belongs to exactly one topic, but requires a conditional independence assumption on word and topic.

- The language model probability now comprises two components $Pr\left(t_k \mid h_q\right)$ and $Pr\left(w_q \mid t_k\right)$, where topic n-gram is taken from the background SLM, i.e., $Pr\left(t_k \mid h_q\right)= Pr_B\left(t_k \mid h_q\right)$ and the topic assignment is adapted as
$$Pr\left(w_q \mid t_k\right)=\left(1-\lambda\right) Pr_A\left(w_q \mid t_k\right)+\lambda\, Pr_B\left(w_q \mid t_k\right)$$

- The main uncertainty in the approach is the granularity required in the topic clustering procedure.

Adaptation Technique - Semantic Knowledge

- Approaches taking advantage of semantic knowledge purpose to exploit not just topic information as before, but the entire semantic fabric of the corpus A .

- We will discuss：
  1. Triggers
  2. Latent Semantic Analysis

Adaptation Technique - Triggers

- A trigger pair is a pair of content words that are semantically related to each other. Trigger pairs can be represented as $w_i \rightarrow w_j$, which means the occurrence of word w_i “triggers” the appearance of word w_j , that is, if appears in a text, it is likely that w_j will come up afterwards.

- If w_i and w_j are single words, possible triggers are more than bigram.

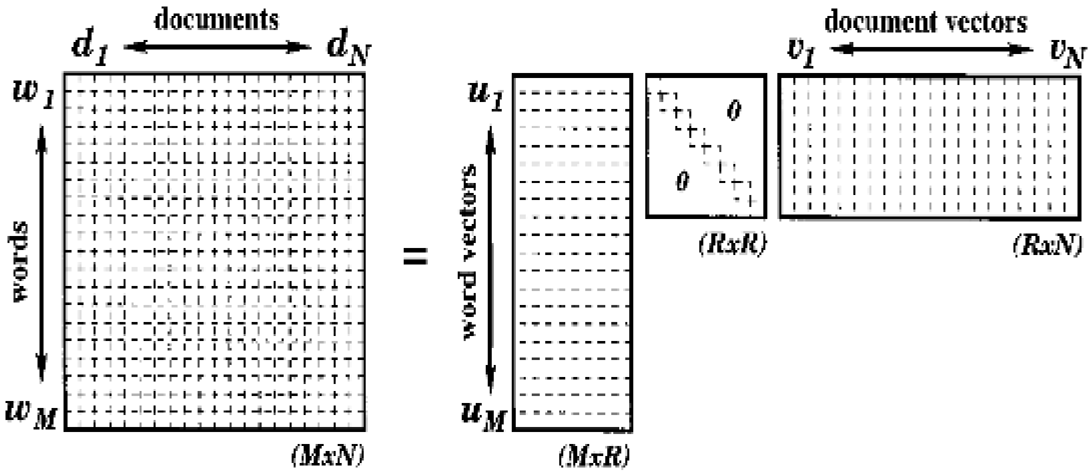
- The method used to extract the trigger pairs is Average Mutual Information which defined：

$$I(w_i : w_j) = P(w_i, w_j) log \frac{P(w_j \mid w_i)}{P(w_j)} + P(w_i, \overline{w}_j) log \frac{P(\overline{w}_j \mid w_i)}{P(\overline{w}_j)}$$

$$+ P(\overline{w}_i, w_j) log \frac{P(w_j \mid \overline{w}_i)}{P(w_j)} + P(\overline{w}_i, \overline{w}_j) log \frac{P(\overline{w}_j \mid \overline{w}_i)}{P(\overline{w}_j)}$$

$$I(w_i : w_j) = log \frac{P(w_i, w_j)}{P(w_i) P(w_j)}$$

Adaptation Technique - LSA



- Let V , $|V|=M$, be the vocabulary and T a training text corpus, comprising N articles (documents) relevant to some domain.

- A suitable expression for the (i, j) cell of W is

$$w_{i,j} = (1 - \varepsilon_i)\frac{c_{i,j}}{n_j}$$

where $c_{i,j}$: *number of times* w_i *occurs in* d_j

n_j : *total number of words present in* d_j

ε_i : *normalized entropy of* w_i *in the corpus* T .

Adaptation Technique - LSA

- The global weighting implied by $1 - \varepsilon_i$ reflects the fact that two words appearing with the same count in d_j do not necessarily convey the same amount of information about the document.

- We denote by $t_i = \sum_j c_{i,j}$ the total number of times w_i occurs in T , the expression for ε_i is easily seen to be：

$$\varepsilon_i = -\frac{1}{logN}\sum_{j=1}^{N}\frac{c_{i,j}}{t_i}log\frac{c_{i,j}}{t_i}$$

- The value of ε_i close to 1 indicates a word distributed across many documents throughout the corpus, while a value of ε_i close to 0 means that the word is present only in a few specific documents.

- The M-by-N (word-document) matrix W resulting from the above feature extraction.

Adaptation Technique - LSA

- When integrating LSA and n-gram together, we start with the definition：

$$Pr(w_q \mid H_{q,1}^{(n+l)}) = Pr(w_q \mid H_{q,1}^{(n)}, H_{q,1}^{(l)})$$

where $H_{q,1}$ denote some suitable admissible history for the particular word w_q, and the superscripts (n) , (l) , and $(n+l)$ refer to the n-gram component $(w_{q,1}, w_{q,2}, \ldots, w_{q,n})$, the LSA component $(d_{q,1})$, where it represent the current document from the first word up to the word w_q , and the integration thereof, respectively.

- This expression can be rewritten as：

$$Pr(w_q \mid H_{q,1}^{(n+l)}) = \frac{Pr(w_q, H_{q,1}^{(l)} \mid H_{q,1}^{(n)})}{\sum_{w_i \in V} Pr(w_i, H_{q,1}^{(l)} \mid H_{q,1}^{(n)})}$$

where the summation in the denominator extends over all words in V . The numerator is seen to be：

$$\begin{aligned} Pr(w_q, H_{q,1}^{(l)} \mid H_{q,1}^{(n)}) &= Pr(w_q \mid H_{q,1}^{(n)}) \cdot Pr(H_{q,1}^{(l)} \mid w_q, H_{q,1}^{(n)}) \\ &= Pr(w_q \mid w_{q,1}, w_{q,2}, \ldots, w_{q,n}) \cdot Pr(d_{q,1} \mid w_q, w_{q,1}, w_{q,2}, \ldots, w_{q,n}) \end{aligned}$$

Adaptation Technique – Syntactic Infrastructure

- Approaches leveraging syntactic knowledge make the implicit assumption that the background and the recognition tasks share a common grammatical infrastructure.

- The background SLM is then used for initial syntactic modeling, and the corpus A to re-estimate the associated parameters.

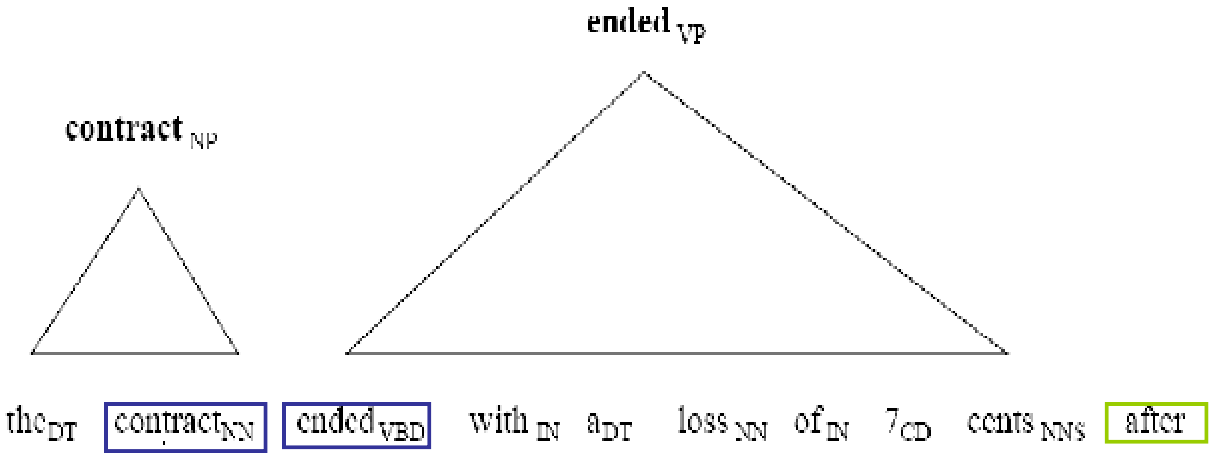
- There are two issue about the approaches：
  1. Structured Language Model
  2. Syntactic Triggers

Adaptation Technique – Structured Model

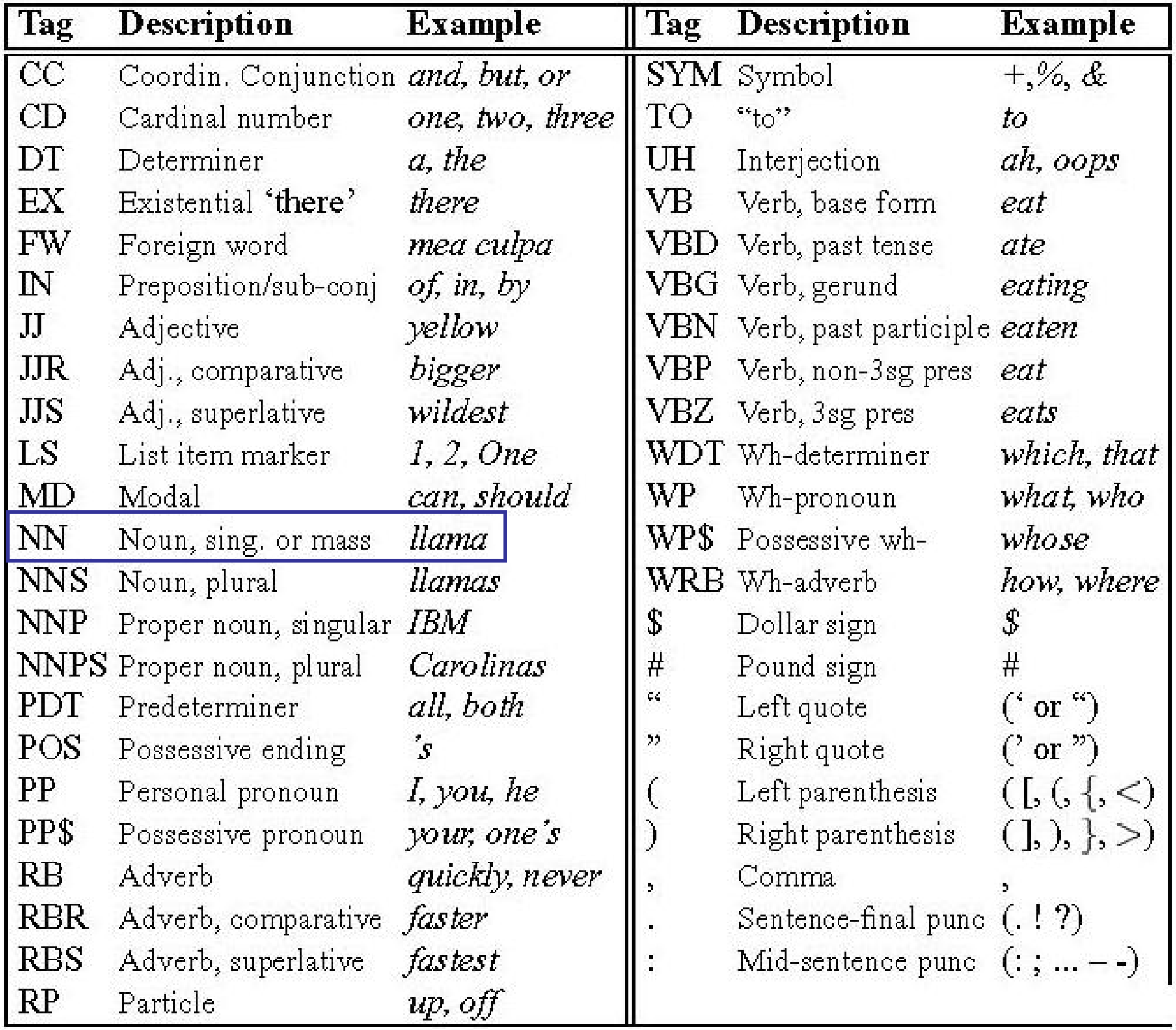
- The main goal of the work is to develop a language model that uses syntactic structure to model long-distance dependencies.

- Structured language modeling takes into account the hierarchical nature of natural language by using syntactic information specifically to determine equivalence classes on the n-gram history.

- For instance, the two head-words preceding the word “after” are “contract” and “ended.”



Adaptation Technique – Structured Model



Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	+,%, &
CD	Cardinal number	*one, two, three*	TO	“to”	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential ‘there’	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WP$	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	$
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	#
PDT	Predeterminer	*all, both*	“	Left quote	(‘ or “)
POS	Possessive ending	*’s*	”	Right quote	(’ or ”)
PP	Personal pronoun	*I, you, he*	(	Left parenthesis	([, (, {, <)
PP$	Possessive pronoun	*your, one’s*	)	Right parenthesis	( ],), }, >)
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	(. ! ?)
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	(: ; ... – -)
RP	Particle	*up, off*			

Adaptation Technique – Structured Model

- This leads to the language model：

  $$Pr\left(w_q \mid \bar{h}_q\right) = \frac{1}{Z\left(\bar{h}_q\right)} \sum_{\{\pi_q\}} Pr\left(w_q \mid \bar{h}_q, \pi_q\right) Pr\left(\bar{h}_q, \pi_q\right)$$

  where $\bar{h}_q$ represents the sentence history so far, $\{\pi_q\}$ denotes the set of all possible parses (or head-words) up to that point, and $Z\left(\bar{h}_q\right)$ ensures appropriate normalization.

- It is expedient to simplify the model by conditioning only on the last $(n-1)$ head-words, denote p_q , as opposed to the entire partial parse and combined with the usual n-gram conditioned on h_q . The final model is given by：

  $$Pr(w_q \mid \bar{h}_q, \pi_q) = Pr(w_q \mid h_q, p_q)$$

Adaptation Technique – Syntactic Triggers

- Structured language models are at the level of the current sentence.

- Another approach extended it by also exploiting syntactic structure contained in previous sentences.

- Although not yet implemented in an adaptation context, this concept may ultimately provide the necessary framework to extend the benefits of structured language modeling to a span greater than that of a sentence.

Adaptation Technique – Multiple Sources

- The approaches exploiting multiple knowledge sources, the corpus is used to extract information about different aspects of the mismatch between training and recognition conditions.

- Two issues ：
  1. Combination Models
  2. Whole Sentence Models

Adaptation Technique – Combination Models

- A popular way to combine knowledge from multiple knowledge sources (such as N-gram, topic tags or syntactic structures) is to use exponential models, because the ME principle guarantees a smooth model that satisfies all these constraints. The method also has the advantage of incorporating an arbitrary number of features while avoiding fragmentation and avoiding data sparseness problems.

- For example, combining n-grams, structured model and topic information, we can get：

  $$Pr\left(w_q \mid \bar{h}_q, \pi_q\right) = Pr(w_q \mid h_q, p_q, t_q)$$

  where t_q corresponds to topic information extracted from the available structured SLM history, $\bar{h}_q$ (i.e., the current sentence so far).

Adaptation Technique – Whole Sentence Models

- All SLM adaptation techniques mentioned so far focus on the adaptation of $Pr(w_q \mid h_q)$, i.e., the probability distribution of a single word.

- One way to improve the drawback is to adopt a “bag-of-features” approach to each sentence, where features are arbitrary computable properties of the entire sentence. This is the case of the whole-sentence exponential model written as

  $$Pr(\sigma)=\frac{Pr_0(\sigma)}{Z}\prod_{k=1}^{K} exp\{\lambda_k I_k(\sigma)\}$$

  where $\sigma = w_1,\dots,w_N$, $Pr_0(\sigma)$ is an initial model estimate, Z is a global constant, and $I_k(\sigma)$ is the feature-selecting indicator functions.

- Note that in this approach, normalization is infeasible, since it involves summation over all possible sentence.

Conclusion

- Language model adaptation refers to the process of exploiting specific, albeit limited, knowledge about recognition task to compensate for any mismatch between training and recognition.
- Model interpolation approaches derive frequency counts from the adaptation corpus and fold them into a well trained SLM.
- Constraint-based methods select promising marginal constraints and other properties of the domain that the background SLM should satisfy, typically within a ME framework.
- Finally, we also can rely on a variety of knowledge sources to appropriately update the semantic and/or syntactic characteristics of the background SLM.

Reference

- Chen, L., Huang, T., 1999. An improved MAP method for language model adaptation. In: Proc. 1999 Euro. Conf. Speech Comm. Technol., Vol. 5. Budapest, Hungary, September 1999, pp. 1923–1926.
- Chen, S.F., Seymore, K., Rosenfeld, R., 1998. Topic adaptation for language modeling using unnormalized exponential models. In: Proc. 1998 Internat. Conf. Acoust. Speech Signal Process., Vol. 2. Seattle, WA, May 1998, pp. 681– 684.
- Federico, M., 1996. Bayesian estimation methods for N-gram language model adaptation. In: Proc. 1996 Internat. Conf. Spoken Language Process., Philadelphia, PA, October 1996, pp. 240–243.
- Federico, M., 1999. Efficient language model adaptation through MDI estimation. In: Proc. 1999 Euro. Conf. Speech Comm. Technol., Vol. 4. Budapest, Hungary, September 1999, pp. 1583–1586.
- Janiszek, D., de Mori, R., Bechet, F., 2001. Data augmentation and language model adaptation. In: Proc. 2001 Internat. Conf. Acoust. Speech Signal Process., Salt Lake City, UT, May 2001.
- Kuhn, R., de Mori, R., 1990. A Cache-based natural language method for speech recognition. IEEE Trans. Pattern Anal. Mach. Intel. PAMI-12 (6), 570–582.
- Carlos TRONCOSO, Tatsuya KAWAHARA, 2006. Trigger-Based Language Model Adaptation for Automatic Transcription of Panel Discussions. Special Section on Statistical Modeling for Speech Processing. IEICE TRANS. INF. & SYST., VOL.E89–D, NO.3 MARCH.
- Ciprian Chelba, Frederick Jelinek. Exploiting Syntactic Structure for Language Modeling.
- Rosenfeld, “A maximum entropy approach to adaptive statistical language model”, 1996.
- Carlos Troncoso et al, “Trigger-Based Language Model Adaptation for Automatic Meeting Transcription”, 2005.