

VOCABULARY AND LANGUAGE MODEL ADAPTATION USING JUST ONE SPEECH FILE

S. Meng^{1,2}, K. Thambiratnam¹, Y. Lin¹, L. Wang^{1,2}, G. Li¹, F. Seide¹

¹Microsoft Research Asia, 5F Beijing Sigma Center, 49 Zhichun Rd., 100080 Beijing, P.R.C.

²Department of Electronic Engineering, Tsinghua University, 100084 Beijing, P.R.C.

{v-shmeng, kit, v-yimlin, v-linfwa, v-gangl, fseide}@microsoft.com

ABSTRACT

This paper investigates unsupervised vocabulary and language model self-adaptation (VLA) from just one speech file using the web as a knowledge source and without prior knowledge of topic or domain beyond optional file metadata. Single-file self adaptation is regularly used for acoustic adaptation, but to date, is rarely used for VLA. The method investigated here uses a first-pass transcript or file metadata to generate web search queries for retrieving texts for adaptation. Various strategies for building queries, retrieving web texts and maximizing out-of-vocabulary (OOV) recovery while constraining vocabulary growth are examined. Significant improvements are demonstrated for transcribing and searching recorded lectures and telephone calls. The proposed method is orthogonal with acoustic adaptation and system combination and integrates well in multi-pass recognition architectures.

Index Terms: Language Model Adaptation, Vocabulary Adaptation, Unsupervised Adaptation, Out-Of-Vocabulary (OOV), Spoken Document Retrieval

1. INTRODUCTION

The general purpose speech recognizer that works for any domain continues to be a dream. In reality, recognizers are built targeting specific domains, typically under the careful guidance of a skilled speech recognition engineer. Yet as CPU cycles become cheaper and digital media becomes more popular, the need for recognizers that work out-of-the-box for any type of content becomes all the more imminent. This paper examines one scenario enabling technology: unsupervised *self-adaptation* of *vocabulary* and *language model* from just *one* speech file using the web as a knowledge source, with no prior knowledge beyond file metadata of topic or domain.

Unsupervised acoustic model (AM) adaptation strategies such as corpus adaptation, self adaptation or online adaptation are well known to be effective in improving recognition accuracy. Today, acoustic adaption is used completely automatically in state-of-the-art speech recognition systems. Unsupervised vocabulary and language model adaptation however is more difficult. Adapting from a first-pass transcript does not allow new words to be discovered and thus limits topic adaptation capability. This limitation is particularly critical for Spoken Document Retrieval (SDR) tasks, where useful query terms like names and jargon are often out-of-vocabulary (OOV). Even supervised domain LM adaptation is problematic since manual transcription of the required amounts of data is costly.

Instead related text documents such as news documents and textbooks [1, 2] are used. For a closed topic set, an ensemble of topic LMs can be trained offline and then selected automatically at recognition time, as described in [3]. Adaptation can also be done online,

for example using a sliding window of the previous days' text news reports to adapt LMs for transcribing future recordings [4, 5].

Such approaches though are cumbersome for general purpose out-of-the-box speech recognition, since defining and sourcing related text sources requires some human intervention. Topic ensemble approaches are also problematic as they enforce a closed topic set that is unsuitable for topically dynamic or heterogeneous content.

An ideal solution would be a method that can be used blindly on a *single* file, without requiring knowledge about the domain or topic of that file. It should not be necessary to aggregate files of a similar topic/domain for batch adaptation, as this is difficult to do for heterogeneous data collections. The ideal approach should not require human effort for defining/collecting related text sources. It should also not require defining a topic set or training topic LMs, as a closed topic set is inappropriate for wildly heterogeneous data sets. What is needed is a method that "just works" out-of-the-box, analogous to MLLR or MAP acoustic adaptation.

The adaptation approach explored here is a step in this direction. Given a single file only, associated metadata (title, speaker name, description) or a first-pass transcript of the audio is used to build queries that are then submitted to a web search engine (or a intranet or domain search engine). Text is then downloaded and used for training a file-specific LM. New words in the downloaded text are ranked and selected to reduce OOV rates while limiting vocabulary growth to, which is important for decoder runtime and footprint.

Similar web-based adaptation has been explored in many previous works [3, 6, 7]. This work primarily differs in that it focuses on both *domain-independent* and *single-file* self-adaptation. It does not enforce domain constraints (e.g. lecture slides in [7] or news texts in [4, 5]) and does not require fixed topics (e.g. [3]). The authors are not aware of any prior work that jointly address all these issues. The presented approach is purposefully a technically straightforward combination of well-known algorithms. The aim is to demonstrate that single-file self-adaptation does result in stable unsupervised adaptation. In addition it shows that this approach is orthogonal with acoustic adaptation and system combination and integrates well in multi-pass recognition architectures.

Section 2 explains the process for sourcing Internet adaptation texts. Vocabulary and LM adaptation are then detailed in section 3 followed by experiments on lecture and telephone corpora in section 4. The paper concludes in section 5.

2. RETRIEVING ADAPTATION WEB TEXTS

The proposed approach for unsupervised Vocabulary and LM Adaptation (VLA) is shown in Figure 1. First a first-pass transcript is segmented into sentence-units (for example using long silences). An automatic Part-of-Speech (POS) tagger is then used to extract Noun

Phrases (NP). Low confidence NPs are thresholded on phrase posterior probability and then adaptation queries are built using one of:

1. **“NP”**: Each NP is submitted *with* phrase quoting.
2. **NP**: Each NP is submitted *without* phrase quoting
3. **NP Prox.**: Noun phrases in proximity (sentences for metadata, segments for recognition transcripts) are combined into a boolean AND query.
4. **Sentence**: All words in a sentence or segmented utterance are issued as a single unquoted query [7].

Each query, q_k , is then sent to an Internet search engine and the top N documents, $\mathbf{D}^k = (d_1^k, \dots, d_N^k)$ are used as adaptation text.

Alternatively, file metadata (includes title, description, speaker name, slide text and other collateral data) can be used for sourcing adaptation queries. If enough NPs can be extracted (at least 30 NPs), then metadata-sourced queries are likely to yield better adaptation, since they would contain OOV terms that may find more relevant adaptation texts. Metadata however is often unavailable or too short and thus adaptation from the first-pass transcript is a useful fall-back.

Longer speech files (e.g. news reports) can contain multiple topics. Ideally, topic segmentation should be applied to obtain topically homogeneous segments for segment VLA. However, experience with topic segmentation, particularly for open topic sets, has found that topic segmentation is not accurate, especially for high word error rates (WERs). Thus, whole files are naively adapted here.

3. VOCABULARY AND LM ADAPTATION

The web-sourced text documents, here termed *adaptation texts*, require text normalization and filtering. This is non-trivial since web text is highly variable in structure and encoding. The filtering used in this work simply removes HTML tags, normalizes acronyms into a standard form, removes spurious punctuation, and normalizes hyphenated words. This normalized text is then used for adapting the vocabulary and language model as described below.

3.1. Vocabulary Adaptation

The purpose of vocabulary adaptation is to augment the recognition dictionary with new words. However, naively adding all newly discovered words in the adaptation text can lead to a very large increase in vocabulary size. In some cases, in excess of 150,000 new words were found in adaptation texts. Adding too many new words leads to an unnecessarily large memory footprint and slower decoding speed. Some of these new words are simply mis-spellings of English words, but heuristically pruning based on counts is ineffective, as these are consistent mis-spellings. Other contributions included multiple realizations of acronyms and abbreviations.

Thus pruning of new words is judicious. The methods used here are derived from the work first presented in [8]. Given an initial vocabulary set, $\mathbf{V}$, the task then is to determine the set of auxiliary words, $\mathbf{V}^*$ that will reduce the OOV rate while constraining the growth $|\mathbf{V}^*|$. The adapted set is given by $\mathbf{V}' = \mathbf{V} \cup \mathbf{V}^*$.

Given the adaptation text $\mathbf{D}^k = (d_1^k, d_2^k, \dots, d_N^k)$ for each query q_k , the word sets and count vectors are extracted to give $W(d_j^k) = (w_1^{j,k}, \dots, w_M^{j,k})$ and $\mathbf{F}(d_j^k) = (f_1^{j,k}, \dots, f_M^{j,k})$ respectively. Then, selection is done using one of:

Greedy: All words in the retrieved document set, $\mathbf{V}^* = W(\mathbf{D}^1) \cup W(\mathbf{D}^2) \cup \dots \cup W(\mathbf{D}^K)$

Unigram: Top- N by count, $r(z) = \sum_k \sum_j f_m^{j,k} |_{w_m^{j,k}=z}$

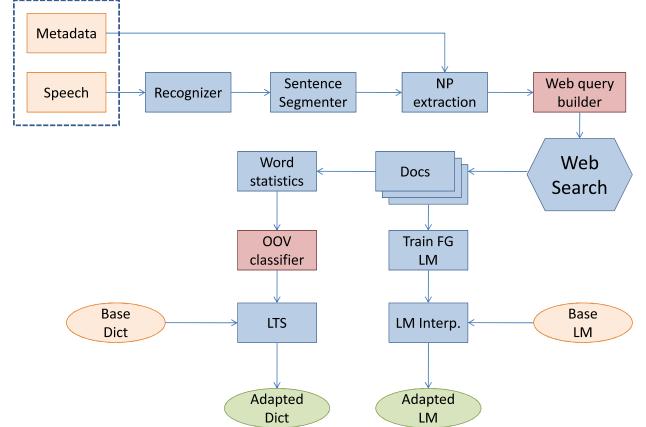


Fig. 1. Unsupervised LM and vocabulary adaptation process

NN: Vocabulary selection is treated as a binary classification task and a neural network is used to classify words as belonging to $\mathbf{V}^*$. The 15-node single hidden layer neural network is trained on a set where manual transcripts of the audio are used to specify $\mathbf{V}^*$. The features used are:

1. Term frequency $TF(w_z) = f_z / \sum_{z'} f_{z'}$
2. Weighted TF: $TFIDF(w_z) = TF(w_z) \times IDF(w_z)$, $IDF(w_z) = \log N_D / N_D(w_z)$, N_D is the document count, and $N_D(w_z)$ is the count of documents with ≥ 1 occurrences of w_z .
3. Tapered TF: $TTF(w_z) = \log(1 + TF(w_z))$. Log tapering reduces sensitivity to large frequencies.
4. Tapered TFIDF: $TTFIDF(w_z) = TTF(w_z) \times IDF(w_z)$
5. The coded part of speech tag for w_z .

The new words are then added to the original dictionary. A letter-to-sound Classification and Regression Tree (CART) is used to generate pronunciations. Special handling is required here as typical letter-to-sound CARTs generate truncated or nonsensical pronunciations when presented with the non-english and nonsensical words that appear in web texts. Truncated pronunciations are easily removed using length heuristics, however nonsensical pronunciations are kept as it is not known how to detect them reliably.

3.2. LM Adaptation

LM adaptation is performed by linear interpolation of a foreground LM (FG-LM) trained on the adaptation text, with a well-trained spontaneous speech background LM. The resulting LM will have a vocabulary that is a superset of the vocabulary selected by vocabulary adaptation. Two strategies were explored to address this mismatch. First, only text sentences with no OOVs relative to the new vocabulary were used for the foreground LM. In this way entire probability mass was used by the adapted LM. Alternatively, all sentences were used for the foreground LM, and then n-grams were discarded without updating the LM probabilities. The second approach led to better WER as it used more data, and thus this approach is used for subsequent experiments.

4. EXPERIMENTS AND RESULTS

Experiments were conducted on the MITWorld technical lecture corpus [9]. Development (Dev) and evaluation (Eval) sets of 14 and 20

Table 1. Results on MITWorld Dev set. VLARec[ref] used the reference instead of automatic transcript.

Configuration	WER	FOM		OOV coverage	
		INV	OOV	select	greedy
Baseline	40.7	67.6	0.00	–	–
VLAMeta	39.3	74.4	40.3	60	77
VLARec[$p = 0.2$]	38.9	74.1	34.9	52	88
VLARec[$p = 0.5$]	39.3	73.7	33.4	50	84
VLARec[$p = 0.8$]	40.0	72.5	29.8	47	77
VLARec[ref]	38.4	77.3	36.5	55	87

lectures, respectively, were used. Another 35 lectures were used for vocabulary selection neural network training. Lecture abstracts with 292 words and 83 NPs on average were used as metadata.

A 72-mixture triphone AM ML-trained on 2000h of Switchboard and Fisher data was used with a base 3-gram LM trained on a mixture of telephone conversations, broadcast news and lectures. A 50k vocabulary was used (5.1% OOV rate on the Dev set).

The LVCSR lattice method in [10] was used for SDR. The SDR evaluation query set was built using NPs from manual transcripts. 6325 queries were selected with 1531 OOV queries. In-Vocabulary (INV) and OOV SDR accuracies were measured using the Figure Of Merit (FOM) defined as the detection/false-alarm curve averaged over [0..10] false-alarms/keyword/hour. Vocabulary selection (Select) and Greedy (Greedy) OOV recovery were also measured.

VLA from metadata and 1st-pass transcript are reported as VLAMeta and VLARec respectively. A default setup of “NP” queries, 100 docs/query, NN vocabulary selection, 15k vocabulary growth and a 0.7 foreground LM interpolation weight was used.

4.1. Transcript vs. Metadata adaptation

Initial experiments are shown in Table 1. The baseline system had a WER of 40.7% and FOMs of 67.6% INV and 0% OOV. VLAMeta achieved a 1.4% abs WER reduction and FOM increases of 6.8% INV and 40.3% OOV. VLARec was evaluated at different NP posterior probability thresholds and $p = 0.2$ gave the best results, with a 1.8% abs WER reduction and FOM increases of 6.5% INV and 34.9% OOV. VLARec outperformed (significant) VLAMeta for transcription as VLARec systems yielded more adaptation text and thus had a better trained LM. However, VLAMeta had better OOV recovery and OOV FOM, since the metadata contained some OOVs.

It is interesting that a low p threshold gave the best VLARec result, as more errors would have been used as queries. This is likely because errors are unlikely to be topically correlated and thus will introduce random rather than systematic noise in the queries. However, the $p = 0.2$ system required downloading large amounts of text, at times in excess of 100Mb. Download size for $p = 0.5$ was 50% less and thus is practically a better choice.

The VLARec[ref] result shows the best achievable WER given perfect transcription (but excluding OOVs for adaptation queries). This was 0.5% abs below the $p = 0.2$ system. The gap here is the effect of recognition errors on VLARec.

4.2. Adaptation text retrieval

Experiments to quantify the effect of different query selection strategies are shown in Table 2. The best (with regards to both WER and SDR) query selection approach was NP Prox., with a WER reduction of 1.5% abs and FOM increases of 7.3% INV and 46.9% OOV. Interestingly gains were higher for NP compared to “NP”. A quoted query should have resulted in more relevant texts, however the 5% higher greedy OOV coverage for NP over “NP” suggests otherwise.

Table 2. WER/SDR for tuning of VLAMeta on MITWorld Dev set.

		FOM		OOV coverage	
Configuration	WER	INV	OOV	select	greedy
<i>Query selection:</i>					
“NP”	39.7	74.0	39.2	59	76
NP	39.4	75.1	44.2	62	82
NP Prox.	39.2	74.9	46.9	67	79
Sentence	40.6	72.9	33.7	49	60
<i>Top-N tuning:</i>					
50	39.6	73.8	39.7	59	72
100	39.3	74.4	40.3	60	77
200	39.1	74.7	42.3	63	82
300	38.9	74.4	41.4	60	86
<i>Vocab classifier:</i>					
unigram	39.9	74.3	37.2	54	77
NN	39.9	74.4	40.3	60	
<i>Vocab growth:</i>					
0	39.9	74.6	0.0	–	77
15k	39.3	74.4	40.3	60	
30k	39.2	74.3	43.8	67	
40k	39.2	74.0	44.6	69	
60k	39.2	74.1	45.3	72	
<i>FG-LM weight:</i>					
0.3	38.4	73.0	36.6	60	77
0.5	38.6	74.0	38.9		
0.7	39.3	74.4	40.3		
1	44.0	72.7	41.9		

A likely explanation is that excluding quotes gave some robustness to incorrect NP phrase parsing during POS tagging. Inspection of NP phrases did indeed reveal some clearly incorrect NP phrase groupings. Sentence queries performed poorly as many queries returned very little adaptation text. This is contrary to the result reported by [7] and is because [7] took sentences from presentation slides that would have had compact information-rich sentences.

The Top-N results in Table 2 show the effect of the number of documents retrieved per query. $N = 200$ gave the best FOM and near best WER. $N = 300$ gave a further 0.2% WER reduction but a moderate loss in FOM. This loss was because high information topical queries typically had fewer results than less informative queries and thus their adaptation texts were overwhelmed.

4.3. Vocabulary and LM adaptation tuning

The effect of the vocabulary classifier, vocabulary growth and LM interpolation weight were also investigated and results are shown in Table 2. Vocabulary selection only yielded significant differences in terms of OOV FOM and OOV coverage, with a 3.1% OOV FOM gain for NN over unigram selection. The lack of improvement for WER was expected since the overall contribution of OOVs to WER is very small. High OOV FOM is important for SDR, and thus NN selection should be used for SDR tasks. Practically though, unigram selection is easier to use as it does not require training, and eliminates any concerns regarding the generalization of the NN to other domains. However, experience to date with other data sets has shown that a MITWorld-trained NN does seem to generalize to domains such as conversations and podcasts.

Adjusting the vocabulary growth is useful for constraining decoder footprint and runtime. The results in Table 2 show that adding more words improves OOV FOM as expected, but with a very minor degradation to INV FOM due to increased word confusion. The 0 growth result indicates the gain from only LM adaptation (no vocabulary change). There is a loss of 0.7% in WER, demonstrating that adding new words significantly reduces WER.

LM interpolation weight also significantly affected results as

Table 3. Tuned systems on MITWorld Dev/Eval sets. α = FG-LM weight.

Configuration	WER	FOM	
		INV	OOV
Baseline	40.7 / 41.4	67.6 / 67.7	0.0
VLAMeta[α = 0.3]	37.9 / 38.3	73.6 / 73.8	45.5 / 44.3
VLAMeta[α = 0.7]	38.9 / 38.9	74.9 / 74.7	49.3 / 47.9
VLARec[α = 0.3]	38.4 / 37.9	72.3 / 73.1	37.5 / 41.9
VLARec[α = 0.7]	39.1 / 38.1	73.5 / 74.3	41.0 / 45.7
Pool VLAMeta[α = 0.3]	38.3 / 40.5	72.4 / 71.2	36.1 / 26.7
Pool VLAMeta[α = 0.7]	39.0 / 39.9	73.7 / 70.8	39.8 / 26.2

shown in Table 2, however the optimal weight depended on the task. WER was lowest for a foreground LM weight of 0.3. The spontaneous speech n-grams in the background LM were important for WER since they constituted the majority of the transcript. In contrast, the best weight for SDR was 0.7 (using a 90/10% INV/OOV mix). High-information content words were more important for SDR, and therefore a greater weight for the foreground topic-matched LM resulted in better FOM.

4.4. Tuned systems

Given the above results, a tuned setup (with respect to accuracy, memory and text download size) was defined. This used NP Prox. queries, top 200 documents/query, NN selection, 40k vocabulary growth, and $p = 0.5$ for VLARec thresholding. Table 3 shows the results on the MITWorld Dev/Eval sets. The WER-tuned $\alpha = 0.3$ systems gave a further 1.4/0.9% WER reduction for VLAMeta/Rec on the Dev set compared to the systems in Table 1. WER reductions of 3.1/3.5% were achieved for the Eval set. The $\alpha = 0.7$ systems were tuned for FOM and gave additional FOM gains of 0.5% INV and 9.0% OOV for VLAMeta and -0.2 INV and 7.6% OOV for VLARec. FOM gains on the Eval set were 7.0% INV and 47.9% OOV for VLAMeta and 6.6% INV and 45.7% OOV for VLARec.

4.5. Single-File vs. Pooled Adaptation

Table 3 also shows results for pooled adaptation. Here the adaptation text of all Dev set files were combined and a single adapted vocabulary and LM was built for recognition. This results in more adaptation text but less lecture topic specificity. Using this on the Dev set demonstrates batched self-adaptation while testing on the Eval set demonstrates offline corpus-style adaptation (though with a comparably small amount of data). Both cases were outperformed by single-file self adaptation. There is sufficient variability across these technical lecture topics to motivate per-file adaptation. Of course pooled adaptation could be improved if more vocabulary growth was allowed but this would be at the expense of decoder footprint.

4.6. Multi-Pass and LM System Combination

State-of-the-art systems use multi-pass recognition with acoustic adaptation and system combination. VLA is an orthogonal component that can easily be inserted into this chain. Furthermore, applying VLAMeta and VLARec enables LM system combination. Traditionally system combination has only been applied to combine AMs or features since multiple LMs are difficult to source.

Table 4 shows results for MLLR acoustic adaptation, VLA and ROVER. VLA WER improvements were mostly complementary with acoustic adaptation. For example, the VLAMeta had reductions of 2.8/3.1% before MLLR and 2.3/2.5% after on MITWorld Dev/Eval. VLAMeta and VLARec were also complementary for system combination demonstrated by the 0.4% reduction for

Table 4. WER results for MLLR acoustic adaptation, VLA and ROVER.

Pass	Configuration	MITWorld		
		Dev	Eval	SwbEval2000
P1	Baseline	40.7	41.4	22.5
P2a	MLLR	35.8	34.8	20.1
P2b	VLAMeta	37.9	38.3	—
P2c	VLARec	38.4	37.9	21.6
-	ROVER[P2b+P2c]	37.5	37.3	—
P2d	MLLR+VLAMeta	33.5	32.3	—
P2e	MLLR+VLARec	33.7	32.0	19.3
-	ROVER[P2d+P2e]	33.1	31.6	—

ROVERing of MLLR+VLAMeta and MLLR+VLARec. An average WER reduction of 3.0% was achieved over MLLR.

VLARec on the Switchboard 2000 Evaluation (SwbEval2000) set reduced WER by 0.8% abs which in relative terms was less than MITWorld. This was because the base LM included the Switchboard training corpus and thus was already well matched. SwbEval2000 did not have sufficient metadata to evaluate VLAMeta.

5. CONCLUSIONS

This paper has investigated using the web for unsupervised vocabulary and language model adaptation (VLA). Adaptation is done on a per-file basis, not per-domain and thus allows blind indexing and transcription of heterogeneous speech collections. The method described is completely unsupervised and allows either first-pass decoding results or supporting metadata to be used for adaptation. The demonstrated gains were significant, reducing WER by 3.1% abs (7.5% rel.) adapting from metadata and 3.5% abs (8.5% rel.) adapting from recognition results for the MITWorld technical lecture corpus. A 0.8% abs (4.0% rel.) was achieved for the Hub5 Switchboard 2000 Evaluation set. The OOV recovery ability was also significant, raising OOV FOM to 47.9% for metadata adaptation and 45.7% for transcript adaptation. Additionally the single-file approach was shown to outperform both batch and offline adaptation. Finally it was shown that VLA integrates well in multi-pass recognition, reducing WER by 3.2% abs (9.2% rel) for MITWorld.

6. REFERENCES

- [1] D. Y. Kim et al., "CU-HTK 2004 Broadcast News Transcription Systems," in *Proc. ICASSP'05*, 2005.
- [2] L. Nguyen et al., "The BBN RT04 English Broadcast News Transcription System," in *Proc. Interspeech'05*, 2005.
- [3] Q. Shi et al., "Search and Classification Based Language Model Adaptation," in *Interspeech-2008*, 2008.
- [4] A. Allauzen and J. L. Gauvain, "Diachronic Vocabulary Adaptation for Broadcast News Transcription," in *Interspeech-2005*, 2005.
- [5] S. Renals, D. Abberley, D. Kirby, and T. Robinson, "Indexing and retrieval of broadcast news," in *Speech Communication*, 2000.
- [6] I. Bulyko, M. Ostendorf, and A. Stolcke, "Getting More Mileage from Web Text Sources for Conversational Speech Language Modeling Using Class-Dependent Mixtures," in *Proc. HLT'03*, 2003.
- [7] C. Munteanu, G. Penn, and R. Baecker, "Web-Based Language Modelling for Automatic Lecture Transcription," in *Interspeech'07*, 2007.
- [8] C. E. Liu, K. Thambiratnam, and F. Seide, "Online vocabulary adaptation," in *Interspeech'07*, 2007.
- [9] J. Glass, T. J. Hazen, L. Hetherington, and C. Wang, "Analysis and Processing of Lecture Audio Data: Preliminary Investigations," in *Proc. HLT'04*, 2006.
- [10] Z. Y. Zhou, P. Yu, C. Chelba, and F. Seide, "Towards Spoken-Documents Retrieval for the Internet: Lattice Indexing For Large-Scale Web-Search Architectures," in *Proc. HLT'06*, 2006.