

Contents

Contents	iii
1 Introduction	2
1.1 Approaches to Machine Translation	2
1.2 Statistical Machine Translation	4
1.2.1 Statistical Machine Translation	4
1.2.2 Mathematical Foundation of SMT	5
1.2.3 Language Models in SMT	6
1.3 Motivation and Approaches	7
1.4 Outline	10
2 Literature Review	11
2.1 N-gram Language Models	11
2.1.1 Basics	11
2.1.2 Smoothing techniques	12
2.1.3 Higher order N-gram Language Models	13
2.2 Suffix Array Language Model	13
2.3 Class based Language Model	14
2.4 Cache Language Model	15
2.5 Skipping Models	16
2.6 Sentence Mixture Models	17
2.7 Content based Adaptation of Language Models	19
2.8 Combination of Language Models	20
2.9 Meta Cache Language Model	21
2.10 Summary	21
3 Data and Implementation	23
3.1 Baseline Language Models	23
3.2 Evaluation corpora	23
3.2.1 Spanish BTEC Travel corpus	23
3.2.2 Japanese BTEC Travel Corpus	24
3.2.3 Darpa TIDES Data Sets	24
3.2.4 Gigaword Xinhua Corpus	25
3.2.5 Xinhua LDC Corpus	25
3.2.6 WebBlog Data	25

3.3	Implementation Details	26
3.3.1	Word Cache Language Model	26
3.3.2	Bi- and Trigram Cache Language Model	26
3.3.3	N-best list Cache Agreement Model	28
3.3.4	Class Based Language Model	29
3.3.5	Sentence Mixture Model	29
3.3.6	Language Model Adaptation	33
4	Experimental Results	35
4.1	Class based Language Model	35
4.1.1	Perplexity	35
4.1.2	Translation	36
4.2	Word Cache Language Model	37
4.2.1	Perplexity	37
4.2.2	N-Best List Rescoring	38
4.3	Bi- and Trigram Cache Language Model	39
4.3.1	Initial Experiments	39
4.3.2	N-best List Rescoring	41
4.3.3	Translation	42
4.3.4	MER Translation	45
4.3.5	Cache Language Model Test Results	46
4.3.6	BTEC Japanese English Translation	47
4.4	N-best List Cache Agreement Model	48
4.4.1	N-best List Rescoring	48
4.5	Sentence Mixture Language Model	50
4.5.1	Training Clustering Duration	50
4.5.2	Using Clustering Duration	51
4.5.3	Perplexity	52
4.5.4	Translation with Monolingual Training Data	54
4.5.5	Translation with Bilingual Training Data	55
4.5.6	Sentence Mixture Model Test Results	57
4.5.7	Combining Sentence Mixture Model and more Data	58
4.6	Language Model Adaptation	59
4.6.1	Perplexity	59
5	Analysis	60
5.1	Class Language Model	60
5.2	Cache Language Model	61
5.3	Sentence Mixture Language Model	62
6	Conclusions and Future Work	66
6.1	Conclusions	66
6.2	Future Work	68
6.2.1	Analysis Toolkit	68
6.2.2	Development Set Adaptation	68

6.2.3 Back Translation Future Prediction	69
A Changes to the SRI Toolkit	70
A.1 CacheLM.h	70
A.2 CacheLM.cc	72
B Changes to the LEMUR Toolkit	78
B.1 Offline Cluster main file	78
B.2 Cluster main file	80
B.3 ClusterDB.cpp	81
C StopWords for Sentence Mixture Model	83
List of Figures	85
List of Tables	87
Bibliography	91