IDIAP COMMUNICATION REPORT

DOMAIN-SPECIFIC LANGUAGE MODEL ADAPTATION: A CASE STUDY

Gwénolé Lecorvé Petr Motlicek John Dines

Idiap-Com-01-2013

Version of JANUARY 18, 2013

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Domain-specific language model adaptation: a case study

Gwénolé Lecorvé, Petr Motlicek, John Dines
Idiap Research Institute

November 14, 2011

Abstract

Domain language model (LM) adaptation consists in re-estimating probabilities of a baseline LM to better match the peculiarities of a given broad topic of interest. To do so, a yet common strategy consists in retrieving adaptation texts from the Web based on a given domain representative seed text. In this report, we extensively study this process by analyzing the impact of numerous parameters. The domain adaptation is carried on a set of videos dealing with business and management. The achieved results mainly show which Web querying strategies perform the best and how significantly the supervision level of the adaptation process impacts the overall performances.

Contents

1 Introduction

Domain adaptation of an large vocabulary continuous speech recognition system seeks to re-estimate the vocabulary and the language model (LM) of a baseline recognition system in order to fit peculiarities of a given domain. The ultimate goal is to improve the quality of ASR transcripts for a given back-end application. The basic idea to do domain adaptation is to use the Web as an open corpus in order to retrieve domain-specific data providing accurate statistics for n-gram re-estimation and containing relevant out-of-vocabulary words (OOVs).

Based on this idea, the process can be split into the following generic steps:

1. **Extract queries** from a given **seed text** supposed to be representative of the considered domain ;
2. Retrieve and clean a given **quantity of Web pages** using a **Web search engine** ;
3. **Filter** the Web pages in order to improve the quality of retrieved texts ;
4. **Build a new lexicon** by integrating the adaptation data into the baseline system ;
5. **Build a new LM** by integrating the adaptation data into the baseline system.

As highlighted in bold, this process comes along with many parameters and strategies to be defined.

This report aims at listing and analyzing the results obtained for different instantiations of the adaptation scheme. More especially, it seeks to highlight the best strategy for LM adaptation, whereas vocabulary adaptation is left aside. After presenting the experimental setup and data, results are given in Section 3 for different Web data retrieval strategies, while filtering of adaptation data is discussed in Section 4.

2 Experimental setup

This section gives some details about the baseline LVCSR system, the spoken documents used as a use case, and the evaluation measures.

2.1 LVCSR system

The recognition system is based on:

- IHM acoustic models
  - first pass RT09 (59d-PLP+39d-HLDA features, MGE training)
  - second pass RT09 (MLP features, VTLN + CMLLR adaptation, SAT+MGE training)

- a 50K words lexicon directly copied from the RT09 system
  - uppercase
  - no hyphens
  - British spelling
  - manual pronuciations (Bob)
- a 4-gram LM
  - trained on the corpora AMI, ICSI, AMI web, 150M CMU+ICSI+NIST web, 525M Fisher web, Chil web, 175M Fisher topics, and 191M conversational web
  - interpolated on RT06seval
  - pruned using entropy-based pruning with a threshold of 4×10^{-9}.

2.2 Domain specific spoken documents

In the experiments, the domain is represented by 59 videos coming from the business school IMD and ranging between a few minutes long up to more than one hour. Though the broad domain is economy, these videos are of various types[1] address specific problems. Moreover, they have been recorded in different acoustic conditions[2] and, while all videos are in English, speakers may have a more or less strong accent. For each video, the first 5 minutes have been manually transcribed. This reference consists in a total amount of $40,000$ words while the length of the full ASR output is about $100,000$ words.

The adaptation work presented in this report does not focus on precise topics but rather seek to specialize a system a broader domain. Adaptation are thus not performed recording per recording but at the level of sets of spoken documents. To do it, the videos are split into a development set of 30 videos, based on which tunings are performed, and an evaluation set of 29 held-out videos. Hence, references for each of these sets are approximatly the same, i.e., about $20,000$ words. As shown by Table 1, the two sets partially address the same topics since, for a video from a given, there is $40\,\%$ of chance to find a video dealing with the same topic in the other set. This is quite logical since IMD's domain of expertise is charactezied by some very strong topics, such as leadership development, business and team management, teaching, etc. In parallel, each set contains around 40 major different speakers where 13 of them are present is both sets.

2.3 Evaluations measures

The impact of domain adaptation is evaluated in terms of perplexity since the report focus on LM adaptation and perplexity is much faster to compute compared to word accuracy. Word accuracies should be reported in a close future for

[1] Faculty teaching and cases, promotion of programs, services and books, conferences and events, interviews, corporate communication, or research.

[2] Various microphones types, reverberation...

	Number of topics	Number of videos
Dev.	8 / 23	13 / 29
Eval.	8 / 22	11 / 28

Table 1: Topical proximity between the development and evaluation sets. Proximities are given in terms of shared topics and of videos addressing these common topics w.r.t. properties of each set.

the most interesting setups. Measures are directly computed on the concatenation of all the references, as opposed to computing the mean of single evaluations for each video. This does not involve any interpretablity problem since all the reference approximatly contain the same number of words. Finally, even if vocabulary adaptation is not directly targeted here, OOV rates are reported when necessary.

3 Data retrieval

As drawn in introduction, adaptation data retrieval consists in retrieving Web pages by submitting queries to a Web search engine. The parameters to be defined are:

- The seed text from which queries are extracted ;
- The strategy to extract/build queries ;
- The search engine ;
- The quantity of adaptation data, be it in number of pages or in number of words.

Before presenting results, possibilities for every parameter are listed in the following subsections.

3.1 Seed text

The seed text used to extract queries should depend on the adaptation scenario. In the most supervised case, one may assume that a reliable seed text is provided by the user. At the opposite, the unsupervised case corresponds to the situation where only some domain-representative spoken documents are provided without any metadata.

Hence, in the frame of the IMD data, 4 seed texts are considered.

1. The reference of the development set. This is refered as `REF_TRUNC`[3]. This can be seen as the most supervised case.
2. Manually collected Web pages. These pages currently correspond to a section of IMD's website, entitled "Tomorrow's challenges", describing hot

[3] `TRUNC` denotes the fact that the reference does not cover the whole videos.

topics in business and management. This data, refered as IMD_TC, represents an amount of 400,000 word, which is 10 times bigger than REF_TRUNC. This situation somehow stands for a more relaxed supervised case.

3. The ASR transcript corresponding to REF_TRUNC. This is refered as ASR_TRUNC and stands for a first unsupervised case. The wordaccuracy of the ASR is 63.3 %.

4. The full ASR transcript of the development set. This is refered as ASR_FULL. Compared to ASR_TRUNC, this seed text provides a broader view of the domain and should thus answer some question concerning the grain to be adopted for seed texts within domain adaptation.

3.2 Query extraction/building

Two strategies can be adopted to build queries based on a seed texts. It may be either possible to build only a few queries judged as very representative of the seed text and to retrieve as mmy pages as possible for these queries. This first approach can be seen as a search in depth. Or it may rather be possible to consider a very big number of queries in order to span most of the seed text peculiarities. This second approach can be seen as a breadth-first search.

In practice, the depth-first search is based here on n-gram frequencies: among all the 3-grams in the seed texts, all of them containing stopwords (prepositions, articles, model verbs...) are discarded and the N most frequent remaining 3-gram are selected. Then, queries are made of 1 to 3 of these selected 3-trigams, resulting in a total amount of about 60 queries. This strategy is refered as FREQ.

The breadth-first strategy comes from a previous work of [Wan and Hain, 2006]. It consists in considering as a query every 3-grams of the seed text which is not directly modeled by the baseline LM (i.e., trigrams whose joint probability computation would require to backoff to lower order n-grams). By default, this strategy, denoted by UNSEEN, can lead to a very large numbers of queries. Hence, two processings are considered to move to a reasonable number of queries. Either unseen trigrams with at least one stopword are discarded or a cutoff value is set of the n-gram counts. The former is refered as UNSEEN_STOP, the latter as UNSEEN_MIN2, 2 meaning that 3-grams appearing less than twice in the seed text are discarded. Furthermore, the number of queries varies a lot according to the seed text used for their extraction. Table 2 reports the number of queries for the various seed text/querying strategy combinations. It clearly appears that the number of queries depends on the size of the text and that UNSEEN_MIN2 tends to consider twice less queries than UNSEEN_STOP on average.

3.3 Web search engine

Google, Bing and Yahoo! block direct URL-based queries. However, they all (used to) propose the user to go through an API. However, Google API in now very limited in terms of queries and maximum number of search results while Yahoo!'s API services have been shut down. We thus use Bing's API to browse

	REF_TRUNC	ASR_TRUNC	ASR_FULL	IMD_TC
UNSEEN	20073	20053	54205	216565
UNSEEN_STOP	921	818	1948	22276
UNSEEN_MIN2	551	241	1118	10303

Table 2: Number of queries according to the seed text and the query extraction strategy.

the Web. Its only restriction is a limit of 7 queries per second, which is quite reasonable.

3.4 Quantity of adaptation data

To be able to compare the different adaptation methods, it is necessary to work with adaptation corpora of the same size. Hence, for all the querying methods, except for UNSEEN, a pool of $10,000$ pages is retrieved such that the contribution of each query is the same. For UNSEEN, the 10 first results are retrieved for each query without any limit on the overal number of pages.

Then, corpora are built based on the Web pages. However, some querying strategies return Web pages with much lower words, probably because of the higher specificity of some queries. Hence, two sizes have been set for adaptation data. A 5 million words corpus can be extracted for every querying strategy, while corpora of 50M words can be only built for the strategies FREQ and UNSEEN.

3.5 Experiments

For each setting, a domain-specific LM is trained based on the adaptation corpus. This LM is interpolated with the single LMs trained on every background corpus. This interpolation consists in minimizing the perplexity of the interpolated LM with respect to a development text. Whereas this strategy lasts quite long and requires a lot of memory, its main advantage is that it can be considered as the most effective method to generate the adapted LM. Hence, results can be seen as upper bounds regarding the expectable improvements from the adaptation data.

By default, this development text is the evaluation text from RT06 but, within domain adaptation, it can be replaced by a more adequate text, for instance, any of the possible seed texts. The lexicon can also be changed. Two lexicons are considered in the experiments: the default lexicon coming from RT09, and a domain-specific lexicon manually built based on IMD_TC. Both lexicons contain $50,000$ words.

The results seek to highlight the main conclusions regarding the different choices to be made along the adaptation process. For clarity, possible values for the most important parameters are summarized in Table 3. All the results are presented for the IMD development and evaluation sets. Notice that the results on the development set are obviously biased since seed texts are derived from this set. Nonetheless, these results are interesting when compared to those of the

Parameter	Value	Short description
Seed text	REF_TRUNC	Partial reference transcript of the development set.
	ASR_TRUNC	Partial ASR transcript of the development set.
	ASR_FULL	Full ASR transcript of the development set.
	IMD_TC	Web pages collected on IMD's website.
Queries	FREQ	Depth-first search based on most frequent trigrams for a total of 60 queries.
	UNSEEN	Breadth-first search based on all the seed text trigrams which are not directly modeled by the baseline language model.
	UNSEEN_STOP	Trigram containing stopwords are discarded.
	UNSEEN_MIN2	Trigram whose frequency is 1 are discarded.
Data size	50M words	For FREQ and UNSEEN strategies only.
	5M words	For all the strategies strategies only.
Vocabulary	Baseline	Vocabulary from RT09.
	Adapted	Manually adapted vocabulary based on IMD_TC.

Table 3: Summary of the possible values for each parameter of the adaptation data retrieval process.

evaluation set because they provide information about the adaptation process robustness.

3.5.1 Interpolation text

To begin with, Table 4 presents the impact of the interpolation text. Mainly, two kinds of texts are compared: the original texts used to built the baseline LM (refered to as *RT06 eval*), and the different possible seed texts. First, it appears that the evaluation set is a bit harder to model. Then we cam also see that, without integrating new training data, using the reference as an interpolation text provides significant improvements while the other results reflect how degraded the other seed texts are. As expected, using the truncated ASR slightly provides less improvements and using the full ASR leads to even higher perplexities. Finally, it is interesting to notice that the manually collected Web pages IMD_TC perform reasonably well though not being directly refering to the spoken documents. To complete these results, other experiments have shown that integrating adaptation data does not lead to any perplexity improvement when using RT06 eval as the interpolation text. Hence, the choice of a topic-specific interpolation text is mandatory but the exact content of this text is not critical.

Since building a reliable text such as a reference a very expensive, it is quite interesting to study how long the interpolation text needs to be to reach stable and optimal results. By downsampling the REF_TRUNC text with different rates, various interpolation text with different sizes have been used to compute various linearly interpolated adapted LMs. The perplexities of these models against the size of the interpolattion text are presented in Figure 1 for the development

Seed	Queries	Size	Interp. text	Lexicon	Dev.	Eval.
Baseline LM			RT06 eval	Baseline	168	173
Baseline LM			REF_TRUNC	Baseline	**150**	**155**
Baseline LM			IMD_TC	Baseline	153	160
Baseline LM			ASR_TRUNC	Baseline	155	160
Baseline LM			ASR_FULL	Baseline	157	162

Table 4: Perplexities for development and evaluation sets using various seed texts to interpolate the background single LMs.

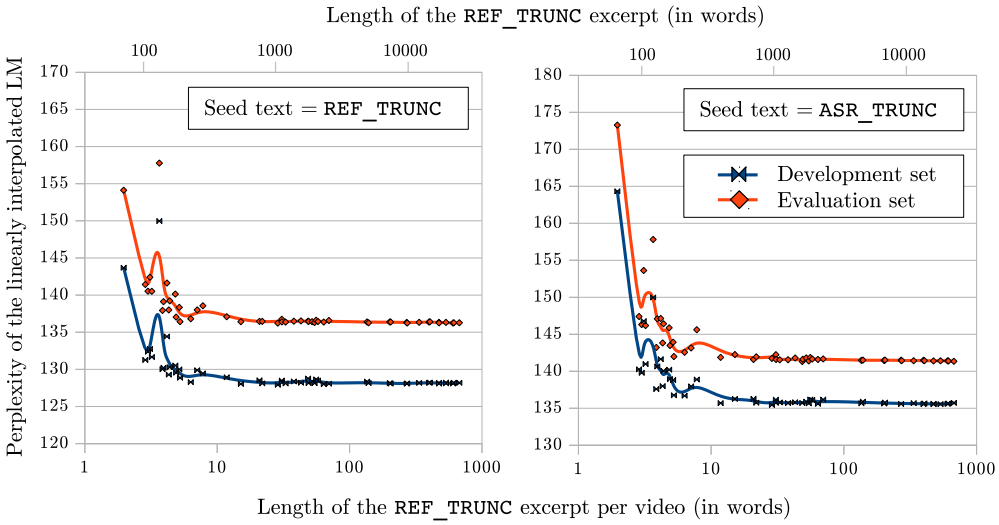

Figure 1: Perplexity versus the size of the linear interpolation text used. Results are presented when using REF_TRUNC (left) or ASR_TRUNC (right) as a seed.

and evaluation sets. On the X-axis, the length of the interpolation text is reported as the overall number of words (top) and as an average number of words for each video of the development set (bottom). It clearly appears that the length of the interpolation text can be drastically decreased from 40,000 words to about 500 words without significantly changing the perplexity of the resulting adapted LM. This is quite low since this represents on average an amount of $15 \sim 20$ words per video from which REF_TRUNC is derived. Under these values, the linear interpolation weight estimation becomes unstable and leads to degraded perplexities. This conclusion is all the more interesting that the same behavior can be observed on the development set and on the evaluation set, be it when using REF_TRUNC or ASR_TRUNC as a seed text.

3.5.2 Seed text

The seed text is an other parameter. It plays a major role within the adaptation data retrieval since it is at the basis of the whole process. Hence, it is important to ensure that the use of a degraded text such as an ASR output does not impact too much the quality of the retrieved texts. To do so, Table 5 presents the first

Seed	Queries	Size	Add. data	Interp. text	Lexicon	Dev.	Eval.
Baseline LM				RT06 eval	Baseline	168	173
REF_TRUNC	FREQ	50M	–	REF_TRUNC	Baseline	**132**	**140**
						(-21%)	**(-21%)**
IMD_TC	FREQ	50M	–	IMD_TC	Baseline	159	172
						(-5%)	(-1%)
ASR_TRUNC	FREQ	50M	–	ASR_TRUNC	Baseline	144	150
						(-14%)	(-13%)
ASR_FULL	FREQ	50M	–	ASR_FULL	Baseline	147	152
						(-13%)	(-12%)
REF_TRUNC	FREQ	50M	–	REF_TRUNC	Baseline	**132**	**140**
IMD_TC	FREQ	50M	–	REF_TRUNC	Baseline	140	148
ASR_TRUNC	FREQ	50M	–	REF_TRUNC	Baseline	139	145
ASR_FULL	FREQ	50M	–	REF_TRUNC	Baseline	140	145
Background texts			+ IMD_TC	REF_TRUNC	Baseline	138	145
REF_TRUNC	FREQ	50M	+ IMD_TC	REF_TRUNC	Baseline	**131**	**139**
IMD_TC	FREQ	50M	+ IMD_TC	REF_TRUNC	Baseline	136	144
ASR_TRUNC	FREQ	50M	+ IMD_TC	REF_TRUNC	Baseline	135	141
ASR_FULL	FREQ	50M	+ IMD_TC	REF_TRUNC	Baseline	136	141

Table 5: Perplexities using various seed texts and the depth-first querying strategy for domain adaptation.

results on perplexity when using adaptation data collected from the depth-first strategy based on different seed texts. Relation variations with the baseline LM are given in brackets. For each setting, the interpolation texts are set as the seed text used to extract queries. This choice is the most realistic one since one may assume that the seed text is usually the only domain-specific text available. We can see that the all the settings lead to perplexity improvement with respect to the baseline LM results of Table 4. The best results are obtained with REF_TRUNC, which is logical since this is the most reliable seed text. Hence, in the remainder of this report, adapted LMs are always trained using REF_TRUNC for the LM interpolation. Then, one can see that the use of the ASR output also leads to perplexity improvement, though gains are considerably smaller. As expected, it appears that using the full transcription degrades the perplexity with respect to using the truncated one but the difference is not significant. Finally, IMD_TC leads to the worse results. This tends us to say that the content of these IMD Web pages does not fit enough the content of the videos to be transcribed in order to drive the retrieval of adaptation data. However, IMD_TC can be used additional adaptation data. As shown by the last series of results, these $400,000$ extra words lead to an absolute perplexity gain of 4 on average, except for REF_TRUNC where the impact is quite null. These can probably be explained the already good quality of the sole corpus retrieved based on REF_TRUNC.

3.5.3 Querying strategy

A comparison of the different querying strategies is given by Table 6. Notice that adaptation corpora for `FREQ` has been shrinked to 5 millions words in order to match the size of corpora built using the two other strategies based on unseen n-grams. Parenthetically, the difference in perplexity with the use of the larger 50M word adaptation corpus is slight compared to the importance of this shrinking. Then, it turns out that the breadth-first strategy leads to significantly better results, be it for the reference or for the ASR. Results for `UNSEEN_MIN2` are slightly worse with respect to `UNSEEN_MIN2` but this may be due to the lower number of queries for thus strategy (see Table 2). Thus, both strategies can be seen as equivalent in terms of perplexity. Results on word error rates might help in distinguishing them.

As a complementary result, one has studied the impact of recognition errors in `ASR_TRUNC` when using this text as a seed along with the breadth-first approach `UNSEEN_STOP`. To do so, misrecognized parts have been removed from `ASR_TRUNC`, remaining unseen trigrams have been listed before retrieving Web pages, and an interpolated LM is trained. Quantitatively, the number of queries falls from 921 to 291. The obtained perplexities (next to last line of Table 6) are the same as when including the recognition errors among the queries. A possible interpretation is that the misrecognized part are the most valuable ones. This seems logical since they precisely represent the unseen n-grams which are the most badly modeled in the baseline LM. However, when only considering these unseen n-grams from the reference of the misrecognized parts for the queries (last line), it appears that the perplexity on the development set decreases compared to the previous setting but that there is not real difference on the test set. Hence, no clear conclusion can be drawn.

3.5.4 Vocabulary

Finally, the use of a topic-specific lexicon has been studied to check if LM adaptation performs better when using a more adequate vocabulary. As a reminder, the topic-specific lexicon has been manually built by mixing words statistics of the background texts with those of the topic-specific 400K word text `IMD_TC`. Results are given in Table 7. The most obvious remark is that perplexity is significantly lower for the baseline LM as well as for adapted ones. Nonetheless, one has to notice that a deeper analysis showed that the adapted lexicon OOV rate is 3 to 4 times higher than the one from the baseline lexicon. Thus, it is unclear if the perplexity decreases come from this discrepancy. Inspite of this, it appears that relative improvements are the same.

As a conclusion for the data retrieval step, the best querying strategy is to use unseen events as queries based on the reference `REF_TRUNC` as a seed text. This corresponds to the supervised case. Nonetheless, the use of ASR transcripts still leads to perplexity improvements. Regarding the text used to estimate the interpolation weights, one has shown that using a reference is clearly better than relying on ASR transcripts or on manually collected domain-specific Web pages. Though generating such a reference is expensive, it also appeared that

Seed	Queries	Size	Interp.	Lexicon	Dev.	Eval.
Baseline LM			RT06 eval	Baseline	168	173
Background texts only			REF_TRUNC	Baseline	150	155
REF_TRUNC	FREQ	5M	REF_TRUNC	Baseline	136	145
REF_TRUNC	UNSEEN_STOP	5M	REF_TRUNC	Baseline	**128**	**136**
					(-24%)	**(-21%)**
REF_TRUNC	UNSEEN_MIN2	5M	REF_TRUNC	Baseline	**128**	139
IMD_TC	FREQ	5M	REF_TRUNC	Baseline	143	151
IMD_TC	UNSEEN_STOP	5M	REF_TRUNC	Baseline	142	147
IMD_TC	UNSEEN_MIN2	5M	REF_TRUNC	Baseline	**140**	**147**
					(-17%)	(-15%)
ASR_TRUNC	FREQ	5M	REF_TRUNC	Baseline	146	150
ASR_TRUNC	UNSEEN_STOP	5M	REF_TRUNC	Baseline	**136**	**141**
					(-19%)	(-18%)
ASR_TRUNC	UNSEEN_MIN2	5M	REF_TRUNC	Baseline	142	147
ASR_TRUNC - errors	UNSEEN_STOP	5M	REF_TRUNC	Baseline	136	142
					(-19%)	(-18%)
reference of errors in ASR_TRUNC	UNSEEN_STOP	5M	REF_TRUNC	Baseline	132	141
					(-21%)	(-18%)

Table 6: Perplexities for development and evaluation sets using various querying strategy.

its length does not need to be too long. A few hundred words is enough. As well, the size of the retrieved adaptation corpus does not seem to be a critical parameter and moving to a topic-specific lexicon does not seem to change the relative improvements of the LM adaptation, though leading in much better absolute results. The next step then consists in investigating the use of filtering strategy to improve the quality of the topic-specifc corpus built on the retrieved Web pages.

4 Data filtering

Data filtering consists in ensuring that the retrieved Web pages effectively fits the target domain. A priori, this appears to be necessary since some queries might be unrelated to the real domain of the spoken documents and thus can possibly lead to irrelevant pages. This should be especially true when the seed is an ASR output.

To filter irrelevant texts, the general idea consists in characterizing the domain before assessing how this characterization suits the different Web pages. To characterize the domain, two methods have been investigated: the TF-IDF model and the Latent Dirichlet Allocation (LDA) model.

Seed	Queries	Size	Interp.	Lexicon	Dev.	Eval.
Baseline LM			RT06 eval	Adapted	149	156
Background texts only			REF_TRUNC	Adapted	136	142
REF_TRUNC	UNSEEN_STOP	5M	REF_TRUNC	Adapted	**114** **(-23%)**	**124** **(-21%)**
ASR_TRUNC	UNSEEN_STOP	5M	REF_TRUNC	Adapted	122 (-18%)	128 (-18%)

Table 7: Perplexities using the adapted lexicon.

REF_TRUNC		ASR_TRUNC	
Score	Word (root)	Score	Word (root)
1.000	I.M.D.	1.000	BUSINESS
0.757	PROGRAM	0.913	COMPANY
0.332	BUSINESS	0.821	PROGRAM
0.327	ORGANISATION	0.713	CHALLENGE
0.268	CHALLENGE	0.551	ORGANISATION
0.266	LEADERSHIP	0.536	LEADERSHIP
0.237	COMPANY	0.495	STRATEGY
0.200	DEBT	0.460	MARKET
0.197	FUTURE	0.457	CHANGE
0.182	STRATEGY	0.440	TALK

Table 8: Partial view of TF-IDF vectors for REF_TRUNC and ASR_TRUNC.

4.1 TF-IDF modeling

The TF-IDF model represents a text as a bag of word where each word is associated with a score depending on its frequency (TF) and its inverse document frequency in a reference collection of texts (IDF) [Salton, 1989]. The higher TF-IDF score, the most discriminant word. Hence, the adaptation domain can be represented by projecting the seed text into the TF-IDF space, which results in a vector of word-score pairs[4].

In our experiments, the reference text collection is made of 3 millions articles from Wikipedia. An outlook of the 10 words with the highest TF-IDF score is given in Table for REF_TRUNC and ASR_TRUNC. At a glance, it clearly appears that the there is a big mismatch between the reference and the ASR due to recognition errors.

Then, similarities can be computed as the cosine between the seed vector and the TF-IDF vector of every retrieved Web page. Low cosine values mean the respective domains of the two documents are completly unrelated. Based on this principle, adaptation corpora are built by selecting the pages with the highest similarities until a size of 5 million words is reached.

[4]Actually, roots are handled instead of words since morphological inflections parasites the term frequency computations.

REF_TRUNC		ASR_TRUNC	
Probability	Word (root)	Probability	Word (root)
8.9×10^{-3}	SYSTEM	9.0×10^{-3}	SYSTEM
7.7×10^{-3}	COMPANY	8.8×10^{-3}	COMPANY
4.1×10^{-3}	ORGANISATION	3.1×10^{-3}	SHOW
3.7×10^{-3}	PROGRAM	3.1×10^{-3}	DEVELOPMENT
3.4×10^{-3}	DEVELOPMENT	3.1×10^{-3}	PROGRAM
3.3×10^{-3}	COMMUNITY	3.0×10^{-3}	ORGANISATION
3.2×10^{-3}	SERVICE	2.9×10^{-3}	PROJECT
3.1×10^{-3}	PROJECT	2.9×10^{-3}	BUSINESS
3.0×10^{-3}	SHOW	2.8×10^{-3}	MARKETING
2.9×10^{-3}	MARKETING	2.6×10^{-3}	SERVICE

Table 9: Top words with the highest LDA probability given REF_TRUNC and ASR_TRUNC.

4.2 Latent Dirichlet allocation modeling

Latent Dirichlet Allocation (LDA) seeks to represent a document as a distribution of a set of a fixed number of latent topic variables, each topic variable being a probability distribution over words [Blei et al., 2003]. These topics are usually broad since their number is small[5]. After infering the various parameters of the model from a training corpus, a new document can be seen as a vector of topic probabilities.

The approach for document similarity computation is the same as with TF-IDF: the seed text and the Web pages are projected into the LDA space, and cosine similarities are computed to build a 5M word adaptation corpus. Documents with highest LDA similarities differ from those based on TF-IDF since LDA does not focus observed words but generalizes the content of a document by backing off onto the broad topic variables. As a comparison, Table 9 shows the most representative words for the topic mixture returned by LDA based on REF_TRUNC and ASR_TRUNC. The probability of a word root w given a document d is computed as follows :

$$P(w|d) = \sum_z P(z|d) \times P(w|z) \ , \tag{1}$$

where z ranges over all latent topic variables. We can see that most important words are more general than those returned by TF-IDF and that the difference between REF_TRUNC and ASR_TRUNC is smaller.

4.3 Experiments

Corpora have been built using the various filtering strategies based on the Web pages retrieved using UNSEEN_STOP. The seed text is the reference or its ASR equivalent while solely reference is used for the final LM interpolation. Table 10

[5]Typically, the number of latent topics ranges between 100 and 1000. In the experiments, it has been set to 100.

Filtering	Seed	Queries	Size	Interp.	Lexicon	Dev.	Eval.
Baseline LM				RT06 eval	Adapted	149	156
Background texts only				REF_TRUNC	Adapted	136	142
Order of retrieval	REF_TRUNC	UNSEEN_STOP	5M	REF_TRUNC	Adapted	114	124
	ASR_TRUNC	UNSEEN_STOP	5M	REF_TRUNC	Adapted	122	128
TF-IDF weighting	REF_TRUNC	UNSEEN_STOP	5M	REF_TRUNC	Adapted	117	127
	ASR_TRUNC	UNSEEN_STOP	5M	REF_TRUNC	Adapted	121	127
LDA model	REF_TRUNC	UNSEEN_STOP	5M	REF_TRUNC	Adapted	117	124
	ASR_TRUNC	UNSEEN_STOP	5M	REF_TRUNC	Adapted	121	127

Table 10: Perplexities for different data filtering models.

presents the perplexity results. Surprisingly, it appears that data filtering has no impact on the resulting adapted LM perplexities. Further experiments should be done to get deeper results, especially on detailed word error rates.

5 Conclusion

In this report, a standard Web-based LM adaptation scheme has been extensively studied by analyzing the influence of every parameter of the adaptation process. A first conclusion is that the best querying solution is to build many queries covering as much as possible the content of a seed text instead of only selecting the few most frequent n-grams. Then, differences have been observed between the supervised case, where a reliable topic-specific seed text is available, and the unsupervised case in which an ASR output is used. The conclusion is that both configurations lead to signigicant perplexity improvements but using the ASR involves much lesser gains. It has also been shown that using an adapted lexicon does not intensify the benefits of adaptation corpora. Finally, experiments show that topic filtering of retrieved data does not produce any effect on perplexity.

All these results should be put in the light of decoding experiments and the first next steps should consist in studying the impact of LM adaptation on various error rates. Then, vocabulary adaptation should also be investigated. Additionally, it could be interesting to revisit some parts of the adaptation scheme, for instance in order to know if recording-centered adaptations, as opposed to dataset-centered, lead to more improvements and if the improvements are stable over spoken documents. Finally, the current process lasts quite long since new LMs are trained from scratch and new ASR outputs are directly generated from the speech signal. Reducing this time while preserving modeling quality gains is thus an other challenge.

References

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

[Salton, 1989] Salton, G. (1989). *Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley Longman Publishing Co., Inc.

[Wan and Hain, 2006] Wan, V. and Hain, T. (2006). Strategies for language model web-data collection. In *Proc. of the Intl Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 1520–6149.