Unsupervised Language Model Adaptation for Automatic Speech Recognition of Broadcast News Using Web 2.0

Tim Schlippe, Lukasz Gren, Ngoc Thang Vu, Tanja Schultz

Cognitive Systems Lab, Karlsruhe Institute of Technology (KIT), Germany
tim.schlippe@kit.edu

Abstract

We improve the automatic speech recognition of broadcast news using paradigms from Web 2.0 to obtain time- and topic-relevant text data for language modeling. We elaborate an unsupervised text collection and decoding strategy that includes crawling appropriate texts from RSS Feeds, complementing it with texts from Twitter, language model and vocabulary adaptation, as well as a 2-pass decoding. The word error rates of the tested French broadcast news shows from Europe 1 are reduced by almost 32% relative with an underlying language model from the *GlobalPhone* project [1] and by almost 4% with an underlying language model from the Quaero project. The tools that we use for the text normalization, the collection of RSS Feeds together with the text on the related websites, a TF-IDF-based topic words extraction, as well as the opportunity for language model interpolation are available in our Rapid Language Adaptation Toolkit [2] [3].

Index Terms: text crawling, language modeling, automatic speech recognition, Web 2.0

1. Introduction

As broadcast news mostly contain the latest developments, new words emerge frequently and different topics get into the focus of attention. To adapt automatic speech recognition (ASR) systems for broadcast news, it is necessary to update the language model (LM) with text data that is in near temporal proximity to the date of the broadcast news show, is part of the same domain, and from the same language. Close temporal and topical proximity of the text data ensures that the words and sentences contained in the news show have a higher probability to fit than using text data from a long time before the show.

We improve the ASR of broadcast news using paradigms from Web 2.0 to obtain time- and topic-relevant text data for language modeling. Web 2.0 is a term coined in 1999 to describe websites that use technology beyond the static pages of earlier websites [4]. A Web 2.0 site may allow users to interact and collaborate with each other in a social media dialogue as creators of user-generated content in a virtual community, in contrast to websites where people are limited to the passive viewing of content. Examples of Web 2.0 include social networking sites, blogs, wikis, video sharing sites, hosted services, web applications, mashups, and folksonomies. In this work we concentrate on the collection of texts from the online social networking and microblogging service Twitter and use information from RSS Feeds that publish frequently updated works such as blog entries and news headlines.

Our motivation is that the Internet community provides there more appropriate texts concerning the latest news faster than on the static web pages. Furthermore texts from older news that do not fit the topic of the show in question can be left out.

The paper is organized as follows: In the next section, we present methods of other researchers to collect and integrate text information for ASR enhancement. Section 3 describes our text collection and decoding strategy. The text and speech data used for our experiments are specified in Section 4. Section 5 presents our experiments, analyses and results. We conclude in Section 6 and give an outlook to future work.

2. Previous Work

Researchers have turned to the World Wide Web as an additional source of training data for language modeling. For example, in [5] the authors achieve significant word error rate (WER) reductions by supplementing training data with text from the Web and filtering it to match the style and topic of the meeting recognition task. Although adding in-domain data is an effective mean of improving LMs [6], adding out-of-domain data is not always successful [7]. To retrieve relevant texts from the Web, search queries are used [5] [8]. Usually search queries are made by extracting characteristic words of every document or web page by calculating a Term-Frequency Inverse-Document-Frequency (TF-IDF) score [9] [10] [11] [12]. [13], [14] and [15] extract topic words from the 1st-pass hypothesis of the show in question and then pass them as a query to a Web search engine. From the retrieved documents, they extract a list of words to adapt the vocabulary for a 2nd-pass decoding. Data from different domains are combined with linear interpolation of N-grams [3] [5]. Furthermore there has been some work on vocabulary adaptation based on word frequency and time relevance [16] [17]. Approaches to obtain time- and topic-relevant text material with machine translation techniques for language modeling in resource-scarce domains and languages are investigated in [18], [19], and [20].

Twitter is an online social networking and microblogging service from Web 2.0 that enables its users to send and read text-based messages of up to 140 characters, known as “Tweets”. Tweets are more real-time than traditional websites and there is a large amount of data available. However, a ristriction is that currently it not possible to get Tweets that are older than 6-8 days with the Twitter REST API[1]. [21] adapts a general LM by interpolating it with an LM trained on normalized TV Tweets and improved ASR accuracy for a voice-enabled social TV application. Another paradigm from Web 2.0 are RSS Feeds. They are small automatically generated XML files, that contain time-stamped URLs of the published updates. RSS Feeds can easily be found on almost all online news websites. [22] uses RSS Feeds to fetch the latest news texts from the Web. [23] subscribe the RSS news Feeds services of six Portuguese news channels for vocabulary and LM daily adaptation for a broadcast news ASR system.

[1] https://dev.twitter.com/docs/api/1.1

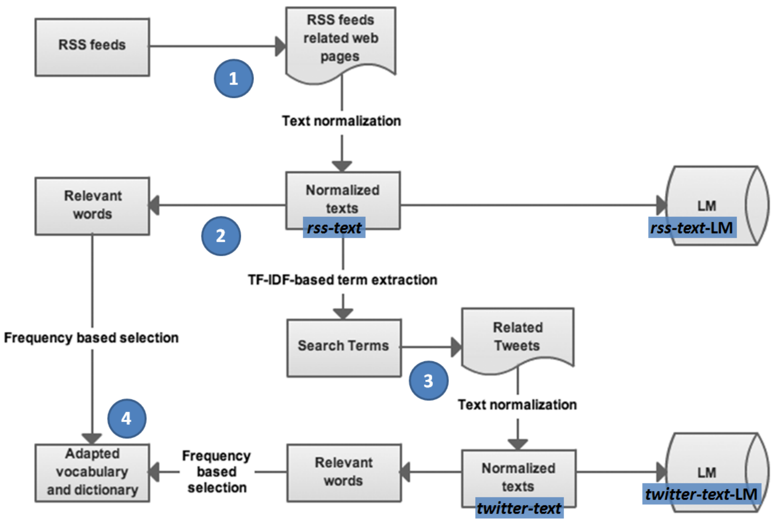

Figure 1: *Strategy.*

In this work we elaborated a strategy to improve LM and ASR quality with time- and topic-relevant text thereby using state-of-the art techniques like TF-IDF-based topic word retrieval, linear LM interpolation, 2-pass decoding, and vocabulary adaptation. Depending on the quality of the underlying generic baseline LM on our test data, we optimize the vocabulary adaptation technique. Our Rapid Language Adaptation Toolkit (RLAT) [2] [3] has the goal to reduce the amount of time and effort involved in building speech processing systems for new languages and domains. We advanced the modules in RLAT for the text normalization, the collection of RSS Feeds together with the text on the related websites, a TF-IDF-based topic words extraction, as well as the opportunity for LM interpolation.

3. Text Collection and Decoding Strategy

Our previous method for collecting large amounts of text data for language modeling was to use the crawler in our RLAT with its recursive crawling implementation [3]. RLAT enables the user to crawl text from a given web page with different link depths. The websites were crawled with a certain link depth, i.e. we captured the content of the given web page, then followed all links of that page to crawl the content of the successor pages (link depth 2) and so forth until we reached the specified link depth. This implementation is good for crawling large amounts of text data. However, it has shortcomings to pick out exclusively text material that is relevant for up-to-date broadcast news shows which we intend to transcribe. To provide text data that fits better to our shows, we extended RLAT with RSS Feeds-based crawling methods.

As shown in Figure 1, our strategy for broadcast news shows starts with the collection of text that is in near temporal proximity to the date of the news show in focus based on URLs in RSS Feeds ①. From this text (*rss-text*), we extract topic bigrams based on a TF-IDF score after text normalization steps as follows ②:

1. Remove stop words in *rss-text* (126 French stop words recommended by `ranks.nl`, a Search Engine Optimization organization[2], worked out to be optimal).
2. Compute the frequency of the bigrams in all downloaded documents where the stop words have been removed.
3. For each bigram, compute the number of documents in which the bigram occurs.
4. The bigrams are scored and sorted in decreasing order according to their TF-IDF score with

   $$score_i = \frac{tf_i}{\sum_j tf_j} ln(\frac{N}{df_j}),$$

   where tf_i is the frequency and df_i is the document frequency of bigram i and N is the total number of downloaded documents.
5. Extract the bigrams with the highest TF-IDF scores (15 bigrams as search queries worked out to be optimal.).

Then we search appropriate Tweets with the resulting bigrams using the Twitter API and normalize them (*twitter-text*) ③. *rss-text* and *twitter-text* are used to build LMs that are interpolated with our generic baseline LM (*base-LM*). To determine optimal interpolation weights, we decode a show in a 1st pass with *base-LM*. Then the combination of weights is adopted that reduces most the perplexity (PPL) on the 1st pass hypothesis. Based on the most frequent words in *rss-text* and *twitter-text*, the vocabulary of the final LM is adapted ④. A 2nd pass decoding with our final LM results in our news show transcription.

4. Corpora and Baseline Language Models

To elaborate and evaluate our text collection and decoding strategy in terms of ASR performance, PPL and out-of-vocabulary (OOV) rate, we downloaded radio broadcasts of the 7 a.m. news from Europe 1[3] in the period from January 2011 to end of February 2012. Each show has a duration of 10-15 minutes. We evaluated our experiments where we included *rss-text* on ten of these episodes. Validating the impact of *twitter-text* was done only on the last 5 shows since we decided to include *twitter-text* in August 2011 and it is not possible to retrieve Tweets older than 6-8 days. Reference transcriptions have been created by a French native speaker. In total, all 10 broadcast news shows contain 691 sentences with 22.5k running words, the last 5 shows 328 sentences with 10.8k running words.

[2] http://www.ranks.nl/stopwords/french.html

[3] http://www.europe1.fr

To investigate the impact of our strategy, we adapted two different baseline 3-gram LMs (*Base*) that have been successfully applied in French ASR but match the domain of our audio data with varying degrees: The French LM from the GlobalPhone corpus [1] (*GP-LM*) and an LM that we used in the Quaero Programme (*Q-LM*). Their average PPLs and OOV rates on the reference transcriptions of all 10 news shows as well as their vocabulary sizes are stated in Table 1.

	GlobalPhone (*G-LM*)	Quaero (*Q-LM*)
Ø PPL	734	205
Ø OOV rate (%)	14.18	1.65
Vocabulary size	22k	170k

Table 1: Quality of our baseline language models.

We collected text data using the information in the RSS Feeds of the 4 French online news websites from *Le Parisien*, *Le Point*, *Le Monde*, and *France24*. All articles which were published up to 5 days before each tested news show were crawled with the RLAT crawler. Totally on average 385k lines from the RSS Feeds-related websites were collected for each show. Further we gathered Tweets that contain 38k lines on average for each show.

5. Experiments

For our experiments we used the acoustic model of our KIT 2010 French Speech-to-Text System [24]. Before the vocabulary adaptation we used the Quaero pronunciation dictionary which has 247k dictionary entries for 170k words. Figure 2 shows the WERs of each Europe 1 show with our *base-LMs*. We see that *Q-LM* performs better than *GP-LM*.

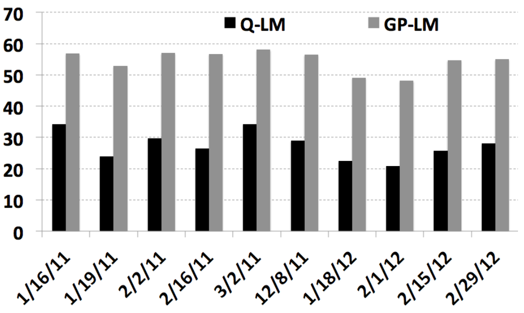

Figure 2: *WERs (%) of the baseline systems*

5.1. 2-Pass Decoding Stategy

With the help of the SRI Language Modeling Toolkit [25], we train individual 3-gram LMs with *rss-text* and *twitter-text* for each show. By interpolating these two Web 2.0-based LMs for the show in question with *base-LM*, we create the LM that we use for the final decoding of the corresponding show (*adapted-LM*). To determine the LM interpolation weights, the following approach is used:

1. Decoding with *base-LM* (*1st-Pass*)
2. Tuning of the interpolation weights for *rss-text*-LM, *twitter-text*-LM and *base-LM* on the *1st-Pass* transcription by minimizing the PPL of the model.
3. Creation of *adapted-LM* from the interpolation of *rss-text*-LM, *twitter-text*-LM and *base-LM* based on these weights
4. Re-decoding with the *adapted-LM* (*2nd-Pass*).

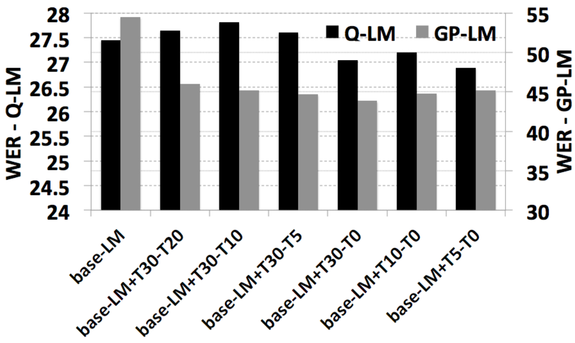

Figure 3: *Average WERs (%) with LMs containing RSS Feeds-based text data from different periods.*

5.2. Time- and Topic-Relevant Text Data from RSS Feeds

We implemented an RSS parser into RLAT, which takes RSS Feeds, extracts the URLs with the publishing date and collects them preserving the time information. Then exclusively the pages corresponding to the listed URLs are crawled. After crawling, HTML tags are removed and the text data is normalized. Our analyses to find the optimal time period for the texts indicate that most relevant texts are from 30 days to the date of the show with *GP-LM* and from 5 days before to the date of the show with *Q-LM*. Figure 3 demonstrates the average WERs with interpolated LMs consisting of RSS Feeds-based text data from different periods of time before the shows. Using text data from less than 5 days to the date of the show decreased the performance. Although for *GP-LM* a *rss-text* collection from 30 days to the date of the show is better than gathering from 5 days before the date of the show, we used only *rss-text* from 5 days before for further experiments with *GP-LM*. The reason is that we had to extract topic words from *rss-text* which are relevant for the search for Tweets and it is not possible to get Tweets that are older than 6-8 days with the Twitter API.

Figure 4 illustrates the average WERs of all 10 tested shows. We see that on average 385k lines of *rss-text* (+*RSS*) for the adaptation of each show improved ASR performance compared to *Q-LM*, while using the same number of lines of randomly selected texts from a recursive crawl of a news website (+*randomText*) decreased the performance. Even 20 million lines of randomly selected texts from traditional recursive crawls did not outperform *rss-text* which indicates its high relevance.

5.3. Time- and Topic-Relevant Text Data from Twitter

From *rss-text*, we extract topic words based on TF-IDF to search relevant French Tweets with the Twitter API in the period from 5 days before to the date of the show. Then we apply the following text normalization steps to the selected Tweets similar to [21]:

1. Remove URLs plus retweet (“RT:”) and mention markers (“@username”),
2. Remove very short Tweets,
3. Remove Tweets being exclusively in uppercase,
4. Remove Tweets containing more than 50% unknown or misspelled words according to French GNU aspell[4],
5. Extend abbreviations.

[4] aspell.net

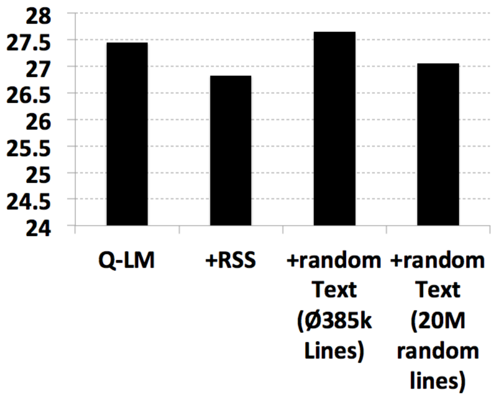

Figure 4: *Average WER (%) with LMs containing RSS Feeds-related text compared to random text data*

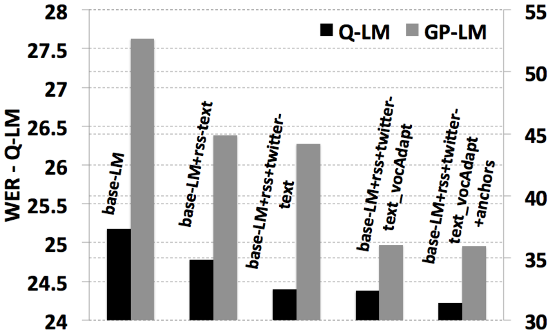

Figure 5: *Results for Q-LM and GP-LM.*

Figure 5 shows that for the last 5 shows on average 1.5% relative WER reduction is achieved by incorporating the *twitter-text*-LM (*Base+RSS+Tweets*) besides the *rss-text*-LM with both underlying *base-LMs*.

5.4. Vocabulary Adaptation

To gain additional performance improvements, we adapt the vocabulary of our LMs (*vocAdapt*) and our decoding dictionary. We experimented with different vocabulary adaptation strategies. The missing French pronunciations were generated with Sequitur G2P, a data-driven Grapheme-to-Phoneme converter developed at RWTH [26], which was trained with the known word-pronunciation pairs from the Quaero dictionary.

For *GP-LM* which has a high OOV rate (13.5%), the following strategy performs best:

1. Generate a list of words present in the concatenation of *rss-text* and *twitter-text* with the corresponding number of occurrences.
2. From this list, we remove all words that are present only once in our text data.
3. The remaining words that are still not present in the search vocabulary are added.

With this strategy on average 19k words are added to the vocabulary for each show. Due to their considerably lower OOV rates, we worked another strategy out for *Q-LM*:

1. Reduce words in the LM to improve the PPL by removing the words with the lowest probability.
2. Remove those words in the decoding dictionary as well.
3. Add the most frequent new words present in the concatenation of *rss-text* and *twitter-text*.

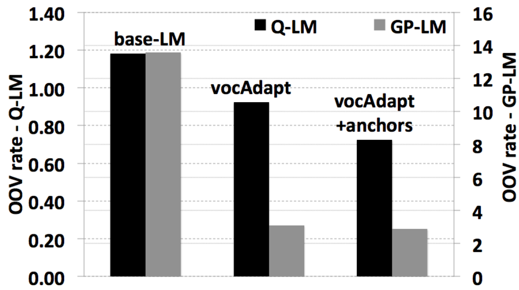

Figure 6: *Average OOV rates (%) for Q-LM and GP-LM before and after vocabulary adaptation*

We experimented with different vocabulary sizes to find a good balance between an increased OOV rate and a lower PPL. Optimal is a new baseline vocabulary with 120k words plus the 1k most frequent words from the concatenation of *rss-text* and *twitter-text*.

Furthermore we manually added the names of the news anchors to the vocabulary as their names were still not present in the adapted vocabulary (*+anchors*). Listening to only one show gives information about the names. The WER reduction with the vocabulary adaptation is shown in Figure 5. The OOV rate decrease is illustrated in Figure 6.

	Q-LM	GP-LM
Adding *rss-text*	1.59	14.77
Adding *twitter-text*	1.53	1.51
Vocabulary adaptation based on *rss-text+twitter-text*	0.08	18.41
Adding names of news anchors	0.66	0.39
Total WER rate improvement	3.81	31.78

Table 2: Relative WER improvement for the last 5 shows with our text collection and decoding strategy

6. Conclusion and Future Work

We have presented a strategy to adapt automatically generic LMs to the several topics that can be encountered during a transcription, especially in broadcast news. We crawled appropriate texts from RSS Feeds, complemented it with texts from Twitter, performed an LM and vocabulary adaptation, as well as a 2-pass decoding. For that we advanced the modules in RLAT for the text normalization, the collection of RSS Feeds together with the text on the related websites, a TF-IDF-based topic words extraction, as well as the opportunity for LM interpolation.

As summarized in Table 2, the WER of five tested French broadcast news shows from Europe 1 are reduced by almost 32% relative with an underlying language model from the *GlobalPhone* project and by almost 4% with an underlying LM from the Quaero project. We have shown the relevance of RSS Feeds-based text and Tweets. Future work may include further paradigms from Web 2.0 such as social networks to obtain time- and topic-relevant text data.

7. Acknowledgements

This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

8. References

[1] T. Schultz, N. T. Vu, and T. Schlippe, "GlobalPhone: A Multilingual Text & Speech Database in 20 Languages," in *The 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013.

[2] A. W. Black and T. Schultz, "Rapid Language Adaptation Tools and Technologies for Multilingual Speech Processing," *The International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008.

[3] N. T. Vu, T. Schlippe, F. Kraus, and T. Schultz, "Rapid Bootstrapping of five Eastern European Languages using the Rapid Language Adaptation Toolkit," in *The 11th Annual Conference of the International Speech Communication Association (Interspeech)*, Makuhari, Japan, 2010.

[4] T. OReilly, "What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software," *Communications & Strategies*, no. 1, p. 17, 2007.

[5] I. Bulyko, M. Ostendorf, and A. Stolcke, "Getting More Mileage from Web Text Sources for Conversational Speech Language Modeling using Class-Dependent Mixtures," in *The 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)*. Association for Computational Linguistics, 2003.

[6] R. Rosenfeld, "Optimizing Lexical and N-Gram Coverage via Judicious Use of Linguistic Data," in *The European Conference on Speech Technology (Eurospeech)*, 1995.

[7] R. Iyer and M. Ostendorf, "Relevance Weighting for Combining Multidomain Data for N-Gram Language Modeling," *Computer Speech & Language*, vol. 13, no. 3, pp. 267–282.

[8] R. Sarikaya, A. Gravano, and Y. Gao, "Rapid Language Model Development using External Resources for New Spoken Dialog Domains," in *The International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, Pennsylvania, USA.

[9] A. Sethy, P. G. Georgiou, and S. Narayanan, "Building Topic Specific Language Models from Webdata using Competitive Models," in *The European Conference on Speech Technology (Eurospeech)*, 2005.

[10] T. Misu and T. Kawahara, "A Bootstrapping Approach for Developing Language Model of New Spoken Dialogue Systems by Selecting Web Texts," in *The Annual Conference of the International Speech Communication Association (Interspeech)*, 2006, pp. 9–12.

[11] G. Lecorve, G. Gravier, and P. Sebillot, "On the Use of Web Resources and Natural Language Processing Techniques to Improve Automatic Speech Recognition Systems," *The Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 2008.

[12] G. Lecorve, G. Gravier, and P.Sebillot, "An Unsupervised Web-based Topic Language Model Adaptation Method," in *The International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2008, pp. 5081–5084.

[13] T. Kemp, "Ein automatisches Indexierungssystem für Fernsehnachrichtensendungen," Ph.D. dissertation, 1999.

[14] H. Yu, T. Tomokiyo, Z. Wang, and A. Waibel, "New Developments In Automatic Meeting Transcription," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, vol. 4, 2000, pp. 310–313.

[15] G. Lecorve, J. Dines, T. Hain, and P. Motlicek, "Supervised and Unsupervised Web-based Language Model Domain Adaptation," in *The 11th Annual Conference of the International Speech Communication Association (Interspeech)*, 2012.

[16] C. Auzanne, J. S. Garofolo, J. G. Fiscus, and W. M. Fisher, "Automatic Language Model Adaptation for Spoken Document Retrieval," in *RIAO 2000 Conference on Content-Based Multimedia Information Access*, 2000.

[17] K. Ohtsuki and L. Nguyen, "Incremental Language Modeling for Automatic Transcription of Broadcast News," *IEICE Transactions on Information and Systems*, vol. 90, no. 2, pp. 526–532, 2007.

[18] S. Khudanpur and W. Kim, "Contemporaneous Text as Side Information in Statistical Language Modeling," *Computer Speech and Language*, vol. 18, no. 2, pp. 143–162, 2004.

[19] S. Kombrink, T. Mikolov, M. Karafiat, and L. Burget, "Improving Language Models for ASR using Translated In-Domain Data," in *The 37th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*. Kyoto, Japan: IEEE, 2012, pp. 4405–4408.

[20] N. T. Vu, D.-C. Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, E.-S. Chng, T. Schultz, and H. Li, "A First Speech Recognition System For Mandarin-English Code-Switch Conversational Speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 4889–4892.

[21] J. Feng and B. Renger, "Language Modeling for Voice-Enabled Social TV Using Tweets," in *The 13th Annual Conference of the International Speech Communication Association (Interspeech 2012)*, Portland, Oregon, USA, 2012.

[22] G. Adam, C. Bouras, and V. Poulopoulos, "Utilizing RSS Feeds for Crawling the Web," in *The Fourth International Conference on Internet and Web Applications and Services (ICIW 2009)*, Venice/Mestre, Italy, 2009, pp. 211–216.

[23] C. A. D. Martins, "Dynamic Language Modeling for European Portuguese," dissertation, Universidade de Aveiro, 2008.

[24] L. Lamel, S. Courcinous, J. Despres, J.-L. Gauvain, Y. Josse, K. Kilgour, F. Kraft, L. V. Bac, H. Ney, M. Nussbaum-Thom, I. Oparin, T. Schlippe, R. Schlüter, T. Schultz, T. F. D. Silva, S. Stüker, M. Sundermeyer, B. Vieru, N. T. Vu, A. Waibel, and C. Woehrling, "Speech Recognition for Machine Translation in Quaero," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT), San Francisco, CA*, 2011.

[25] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit," in *The International Conference on Spoken Language Processing*, vol. 2, 2002, pp. 901–904.

[26] M. Bisani and H. Ney, "Joint-Sequence Models for Grapheme-to-Phoneme Conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.