Karlsruhe Institute of Technology

Cognitive Systems Lab

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

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

Interspeech 2013 – The 14th Annual Conference of the International Speech Communication Association
Lyon, France

-6.049973	Ale	-0.1058635	-2.297622	vaut être bien
-5.729937	Alea	-0.3714143	-2.297622	vaut être blanc
-5.329903	Alec	-0.4949455	-2.297622	vaut être bon
-7.536149	Alecos		-2.013652	vaut être couché
-7.18102	Aleen	-0.3434374	-2.297622	vaut être dans
-6.342624	Alegre	-0.06400765	-2.297622	vaut être dedans
-6.584819	Alegria	-0.04039202	-1.626278	vaut être en
-6.447053	Aleida	-0.1157132	-2.297622	vaut être fasciste
-6.074299	Alejandra	-0.2236585	-1.721409	vaut être le
-5.236938	Alejandro	-0.5320179	-2.297622	vaut être mal
-6.221855	Alejo	-0.4468886	-2.297622	vaut être mort
-7.536149	Alekna		-2.297622	vaut être méprisé
-6.674542	Alekos	-0.1717247	-2.013652	vaut être prudent
-6.307826	Alekperov		-2.132829	vaut être prévenu
-7.18102	Aleksa	-0.1618669	-2.297622	vaut être prévoyant
-6.221855	Aleksandar	-0.5343959	-1.398776	vaut être riche
-6.100069	Aleksander	-0.3111594	-1.585534	vaut être seul
-5.861119	Aleksandr	-0.2324369	-2.013652	vaut être seule
-6.342624	Aleksandra	-0.4656605	-2.132829	vaut être sourd

Outline

1. Motivation and Introduction
2. Text Collection and Decoding Strategy
3. Corpora and Baseline Language Models
4. Experiments
    1. Time- and Topic-Relevant Text Data from RSS Feeds
    2. Time- and Topic-Relevant Text Data from Twitter
    3. Vocabulary Adaptation
5. Conclusion and Future Work

Motivation (1)

- 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) for broadcast news
  - update language model (LM) and pronunciation dictionary

Motivation (2)

- Using paradigms from Web 2.0 *(Oreilly, 2007)* to obtain time- and topic relevant data
  - Internet community provides more appropriate texts concerning the latest news faster than on the static web pages
  - Texts from older news that do not fit the topic of the show in question can be left out
  - Examples:
    - Social networking sites
    - Blogs
    - Web applications

Introduction (1)

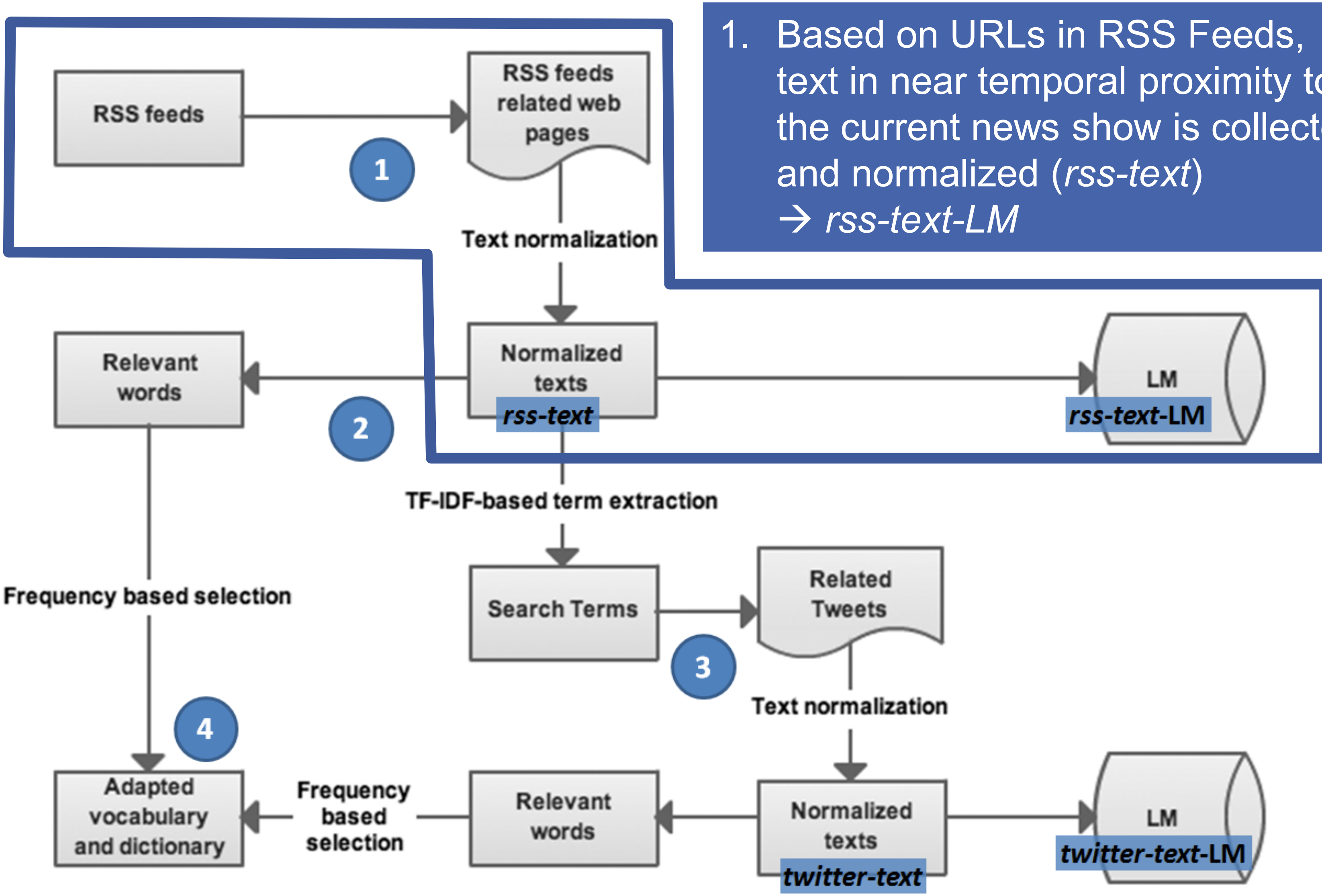
- RSS Feeds
  - Small automatically generated XML files containing time-stamped URLs of the published updates
  - Can easily be found on almost all online news websites
  - Possibility to get data fitting to a certain time interval
- Twitter
  - Enables its users to send and read text-based messages of up to 140 characters (Tweets)
  - Tweets more real-time than traditional websites and a large amount of text data available
  - Restriction: Not possible to get Tweets that are older than 6-8 days with Twitter REST API

Introduction (2)

- Researchers have used WWW as an additional source of training data for language modeling
- Initial works to use Tweets and RSS Feed services
  (Feng and Renger, 2012) (Martins, 2008)

- Our contribution
  - Strategy to enrich the pronunciation dictionary and improve LM with time- and topic-relevant text thereby using state-of-the art techniques
  - Modules for this strategy are provided in our Rapid Language Adaptation Toolkit (RLAT) *(Vu et al., 2010)*

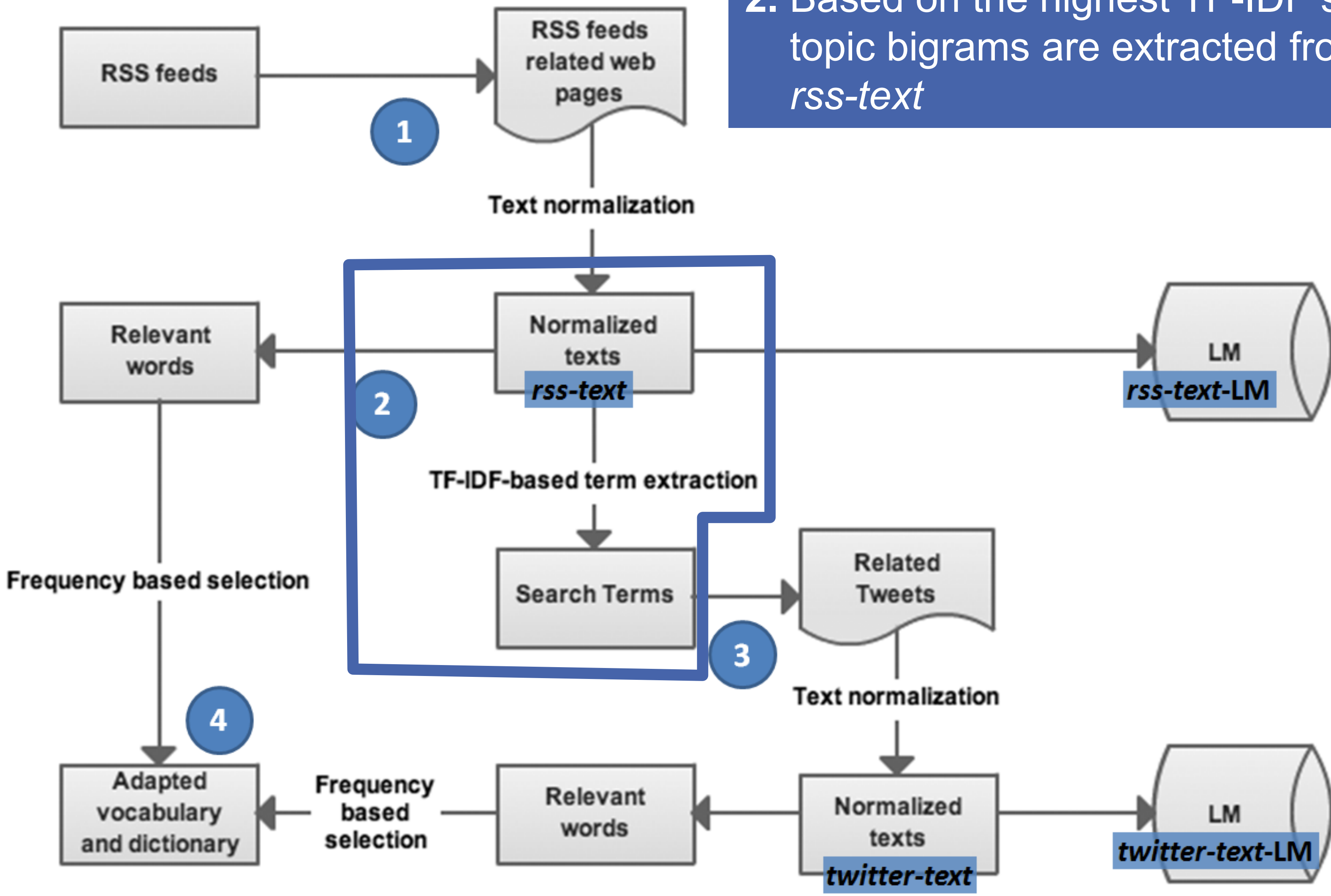
Text Collection and Decoding Strategy (1)



1. Based on URLs in RSS Feeds, text in near temporal proximity to the current news show is collected and normalized (*rss-text*)
   → *rss-text-LM*

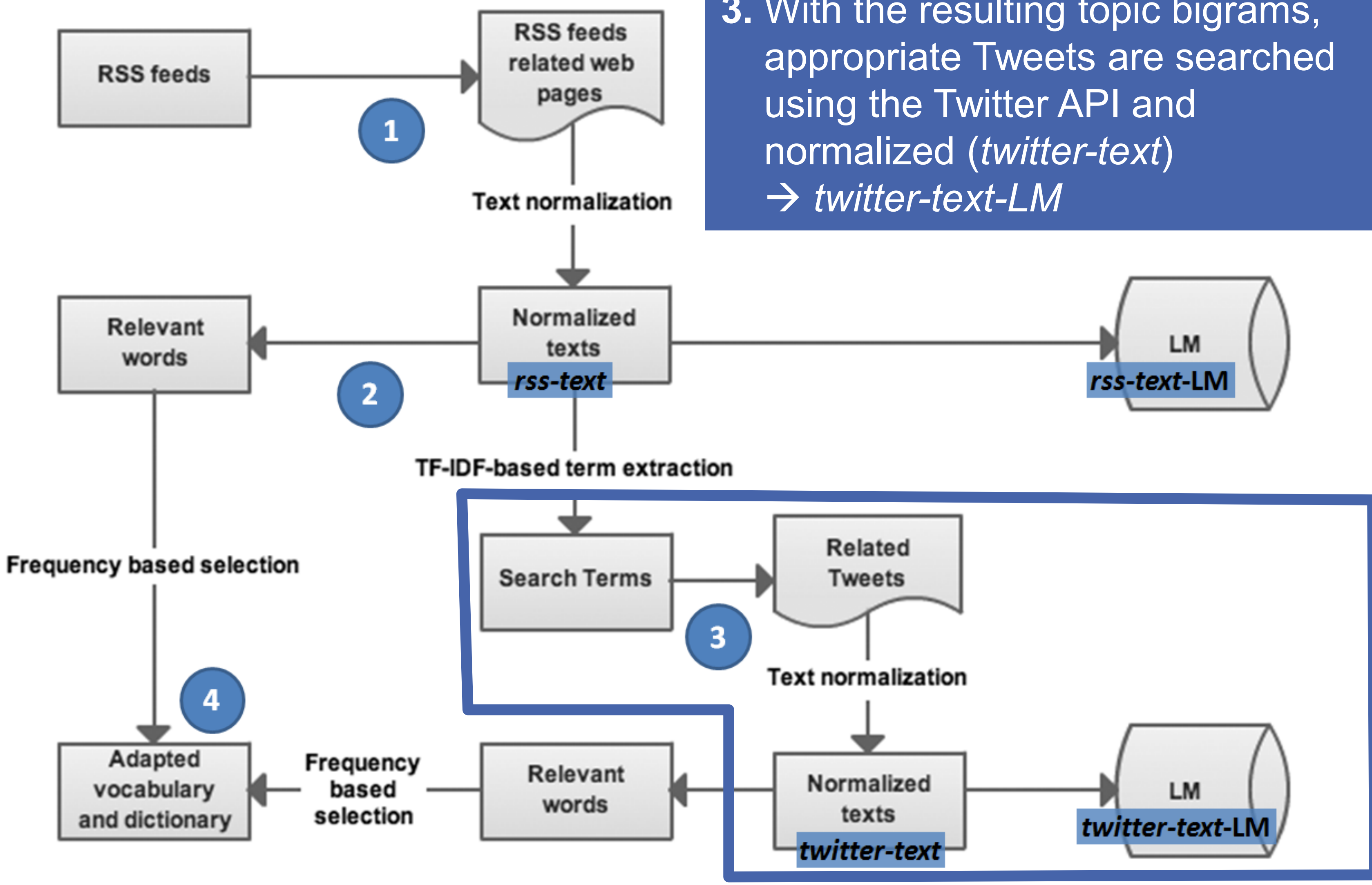
Text Collection and Decoding Strategy (2)

2. Based on the highest TF-IDF score, topic bigrams are extracted from *rss-text*



Text Collection and Decoding Strategy (3)

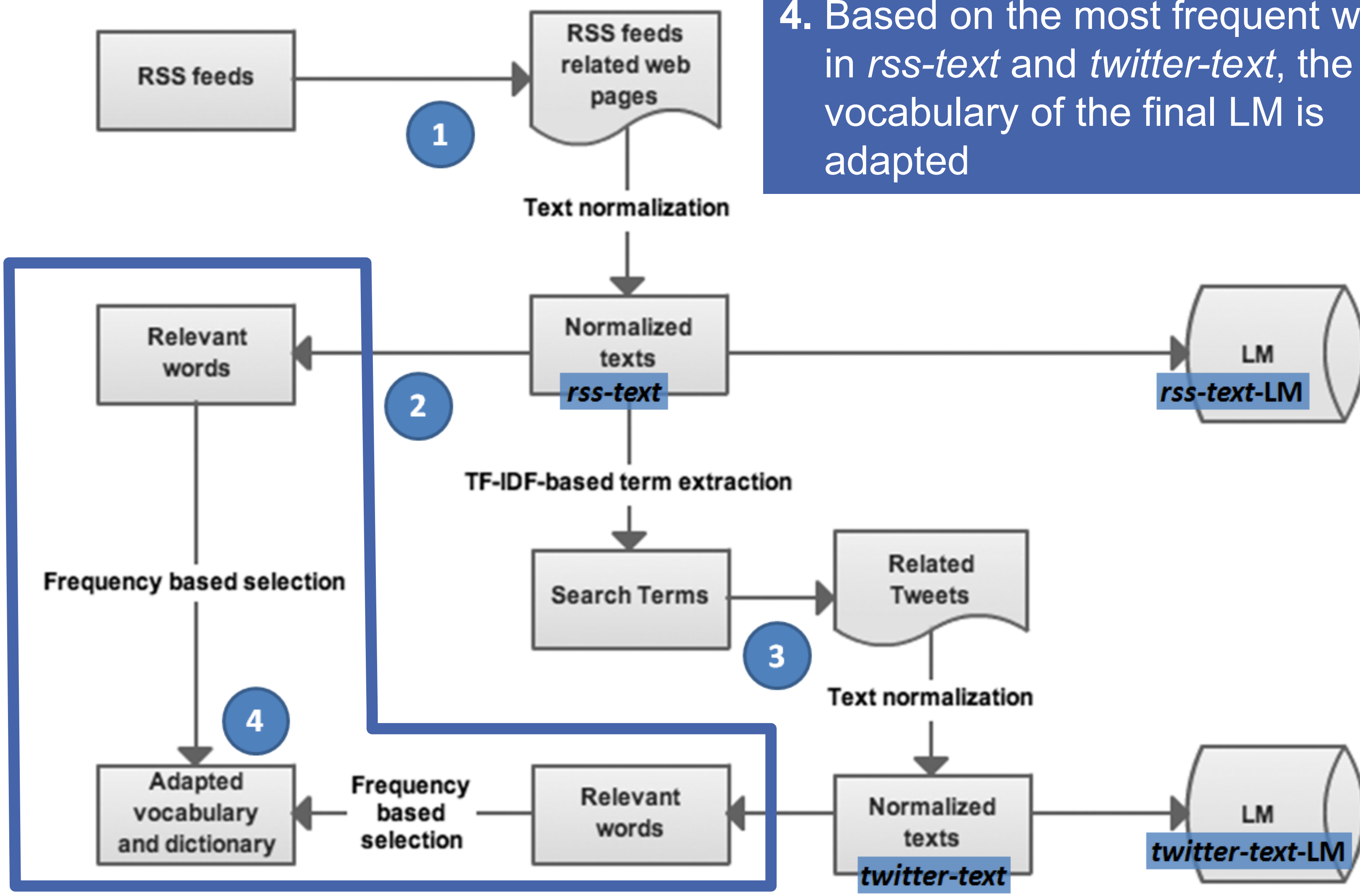
Cognitive Systems Lab



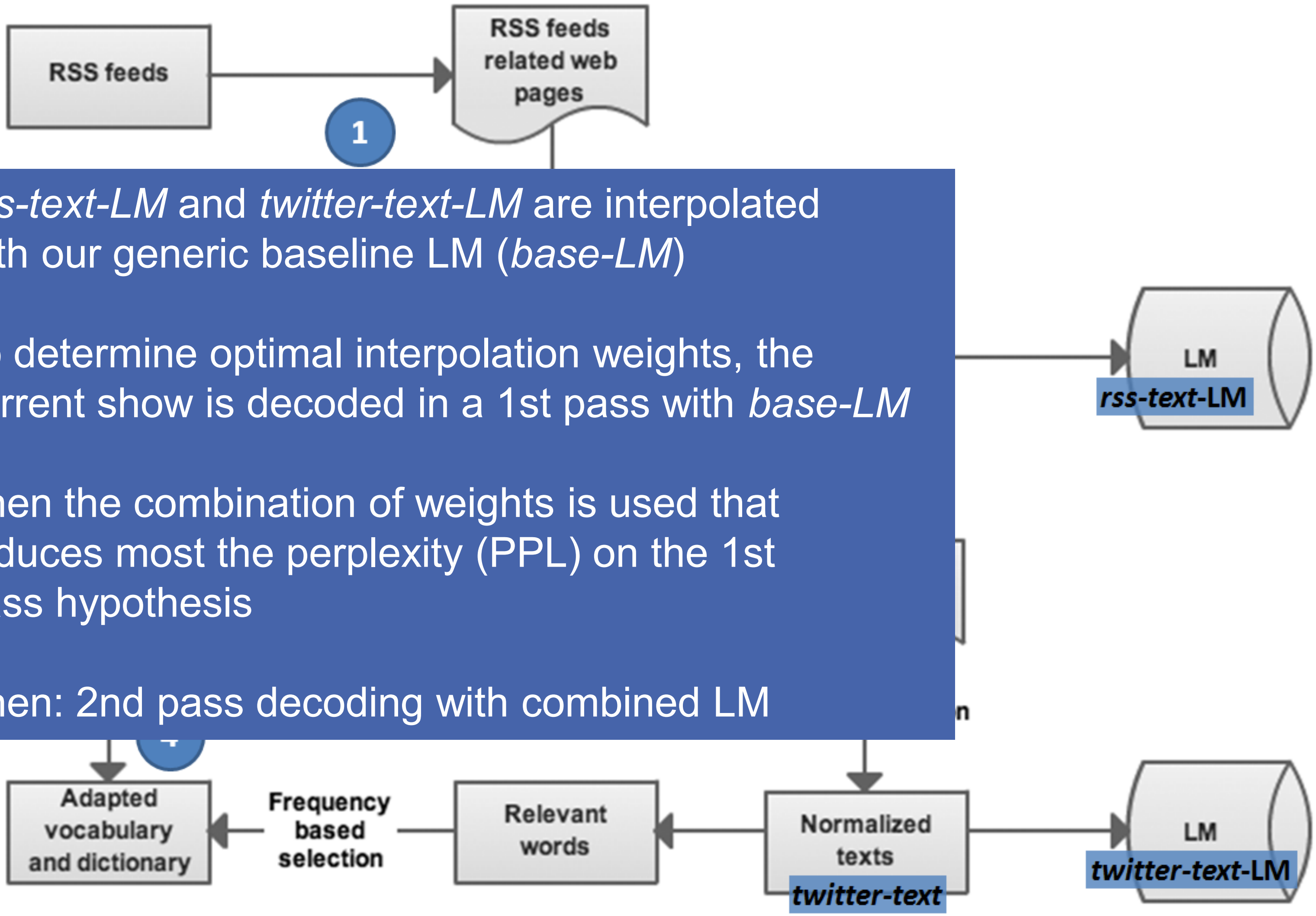
3. With the resulting topic bigrams, appropriate Tweets are searched using the Twitter API and normalized (*twitter-text*)
→ *twitter-text-LM*

Text Collection and Decoding Strategy (4)

4. Based on the most frequent words in *rss-text* and *twitter-text*, the vocabulary of the final LM is adapted



Text Collection and Decoding Strategy (5)



- *rss-text-LM* and *twitter-text-LM* are interpolated with our generic baseline LM (*base-LM*)
- To determine optimal interpolation weights, the current show is decoded in a 1st pass with *base-LM*
- Then the combination of weights is used that reduces most the perplexity (PPL) on the 1st pass hypothesis
- Then: 2nd pass decoding with combined LM

Corpora and Baseline LMs (1)

- Radio broadcasts of the 7 a.m. news from Europe 1
  - Each show 10-15 minutes (French)
  - *rss-text-LM* experiments evaluated on 10 shows
    - 691 sentences with 22.5k running words spoken
  - *twitter-text-LM* experiments evaluated on 5 shows
    - 328 sentences with 10.8k running words spoken
- Subscribed the RSS Feeds services of Le Parisien, Le Monde, France24, Le Point

Corpora and Baseline LMs (2)

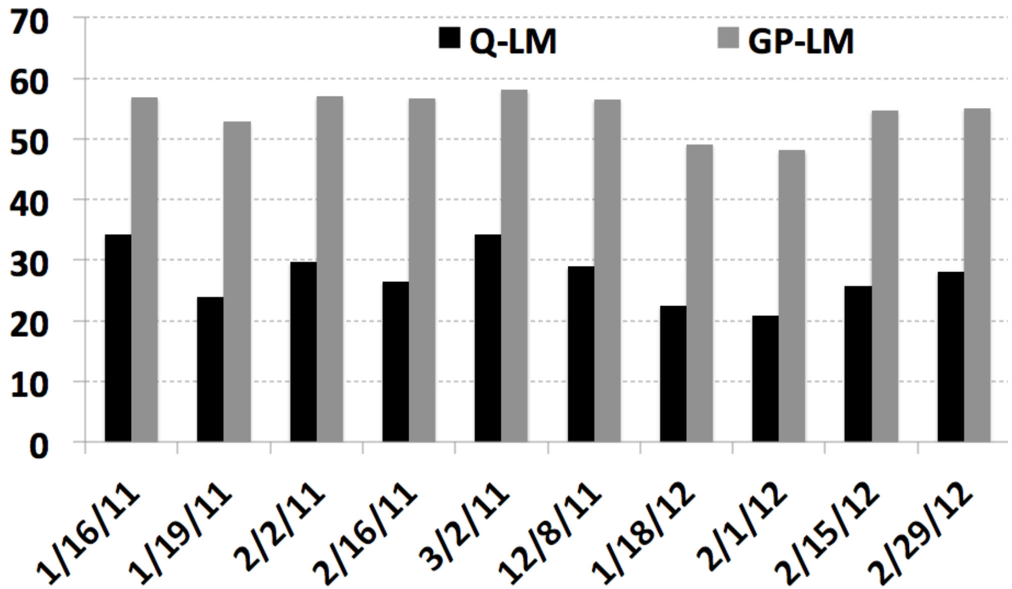
- Strategy analyzed with 2 different baseline 3-gram LMs of different quality (*base-LM*)
  - *GP-LM*: French LM from the GlobalPhone corpus
  - *Q-LM*: French LM that we used in the Quaero project

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

Quality of our baseline language models on the reference transcriptions of all 10 news shows

Experiments

- ASR system
  - Acoustic model of our KIT Quaero 2010 French Speech-to-Text System *(Lamel et al., 2011)*
  - Before vocabulary adaptation: KIT Quaero pronunciation dictionary (247k dictionary entries for 170k words)

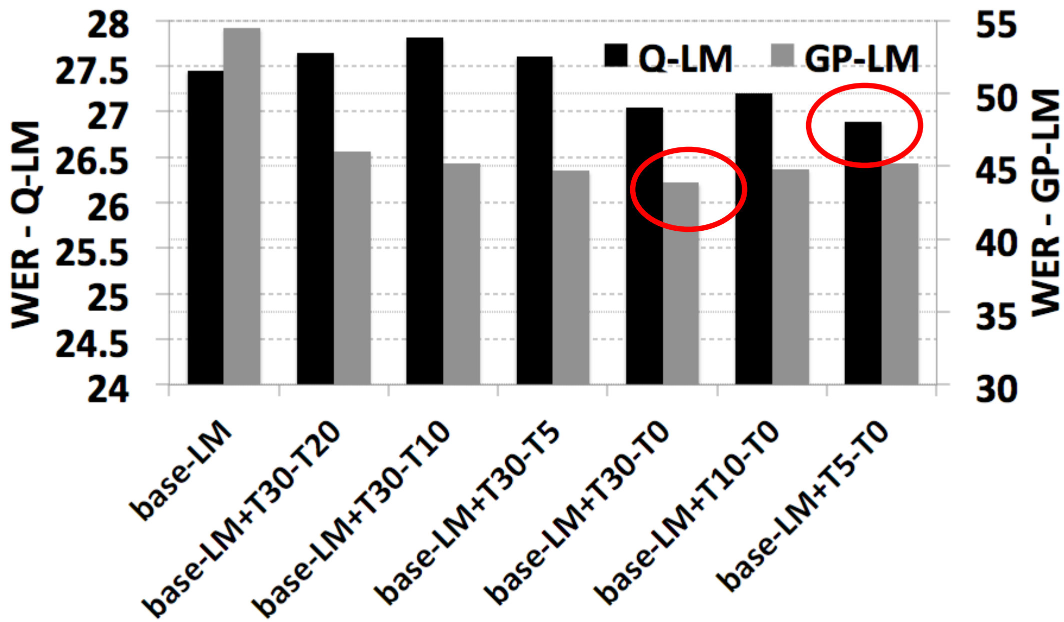


Word error rates (WERs) (%) of our baseline systems

→*Q-LM*: 27.45%
→*GP-LM*: 54.48%

Experiments – Data from RSS Feeds (1)

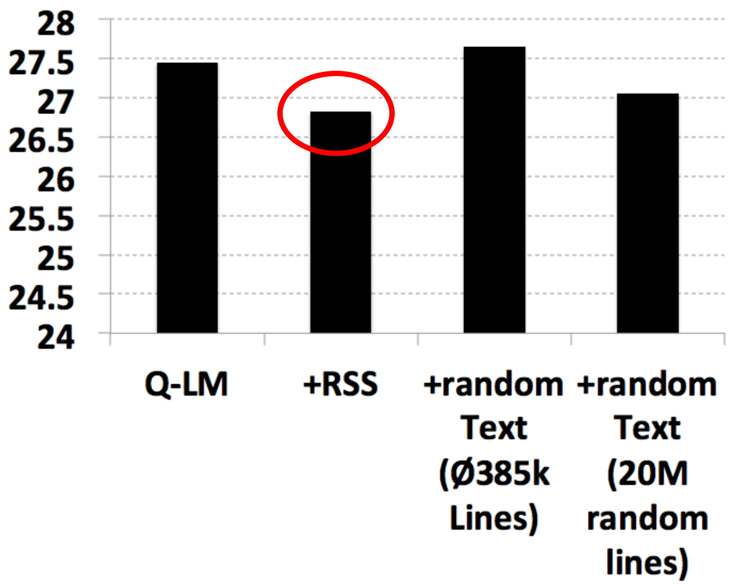
- From which period is *rss-text* optimal?
  - Analyze *rss-text* from different periods



WERs (%) of all shows with LMs containing RSS Feeds-based text data from different periods

Experiments – Data from RSS Feeds (2)

- Is *rss-text* really more relevant than other text?
  - … of the same amount (Ø 385k lines (for each show))?
  - … of a larger amount (e.g. 20M lines)?



WER (%) with LMs containing RSS Feeds-related text compared to random text data

Experiments – Data from Twitter

- From *rss-text*, extract topic words based on TF-IDF
- With topic words, search relevant French Tweets with the Twitter API (in the period from 5 days before to the date of the show)
- Ø38k lines for each show

	Q-LM	GP-LM
Adding *rss-text*	1.59	14.77
Adding *twitter-text*	1.53	1.51

Relative WER improvement for the last 5 shows

Experiments – Vocabulary Adaptation

- Best strategy with *GP-LM*:
  - Include daily on average 19k most frequent words from *rss-text* and *twitter-text*
  - ➔ OOV rate: 13.5% → 3%
  - ➔ WER: 44.22% → 36.08% (18.41% relative)
- Best strategy with *Q-LM*:
  - Remove words with the lowest probability → 120k
  - Include daily on average 1k most frequent words from *rss-text* and *twitter-text*
  - ➔ OOV rate: 1.2% → 0.3%,
  - ➔ WER: 24.40% → 24.38% (0.08% relative)

Overview

	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

Relative WER improvement

- *GP-LM*: 52.68 → 35.94
- *Q-LM*: 25.18 → 24.22

Conclusion and Future Work

- We proposed an automatic strategy to adapt generic LMs and the search vocabulary to the several topics for ASR
- Showed relevance of RSS Feeds and Tweets
- Embedded modules for the strategy into RLAT
- Future work may include further paradigms from Web 2.0 such as social networks or Web 3.0 (Semantic Web) to obtain time- and topic-relevant text data

Merci!

References (1)

[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.

References (2)

[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.