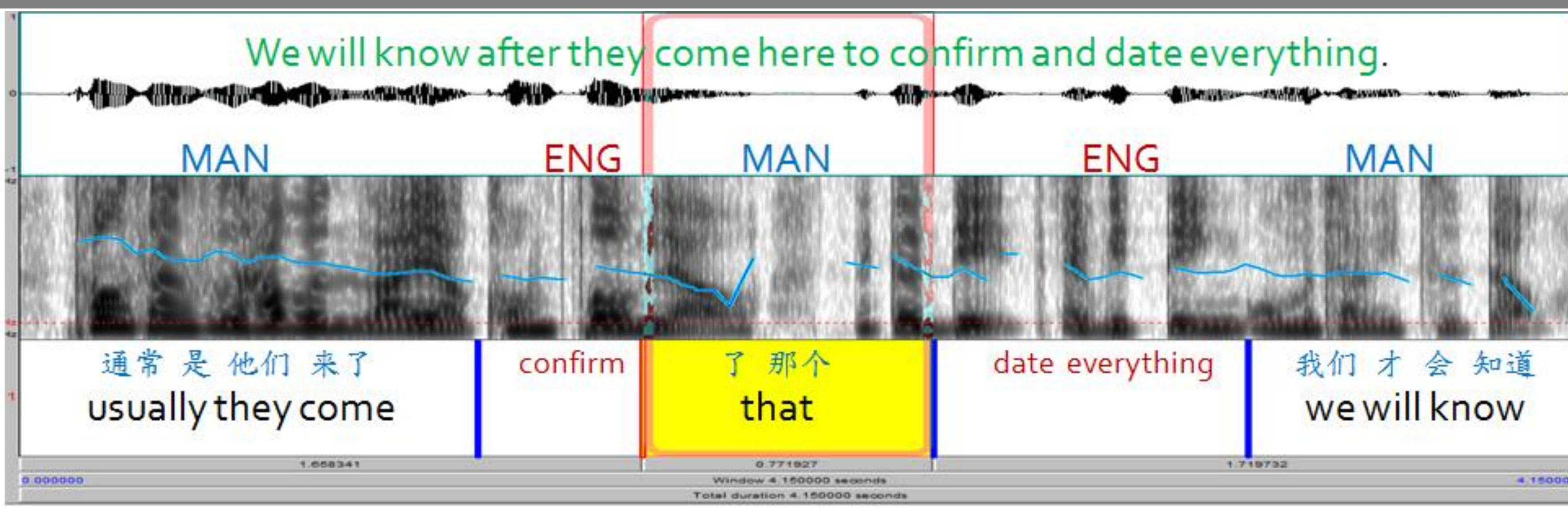
Features for Factored Language Models for Code-Switching Speech

Heike Adel, Katrin Kirchhoff, Dominic Telaar, Ngoc Thang Vu, Tim Schlippe, Tanja Schultz

SLTU 2014 – 4th Workshop on Spoken Language Technologies for Under-resourced Languages
St. Petersburg, Russia

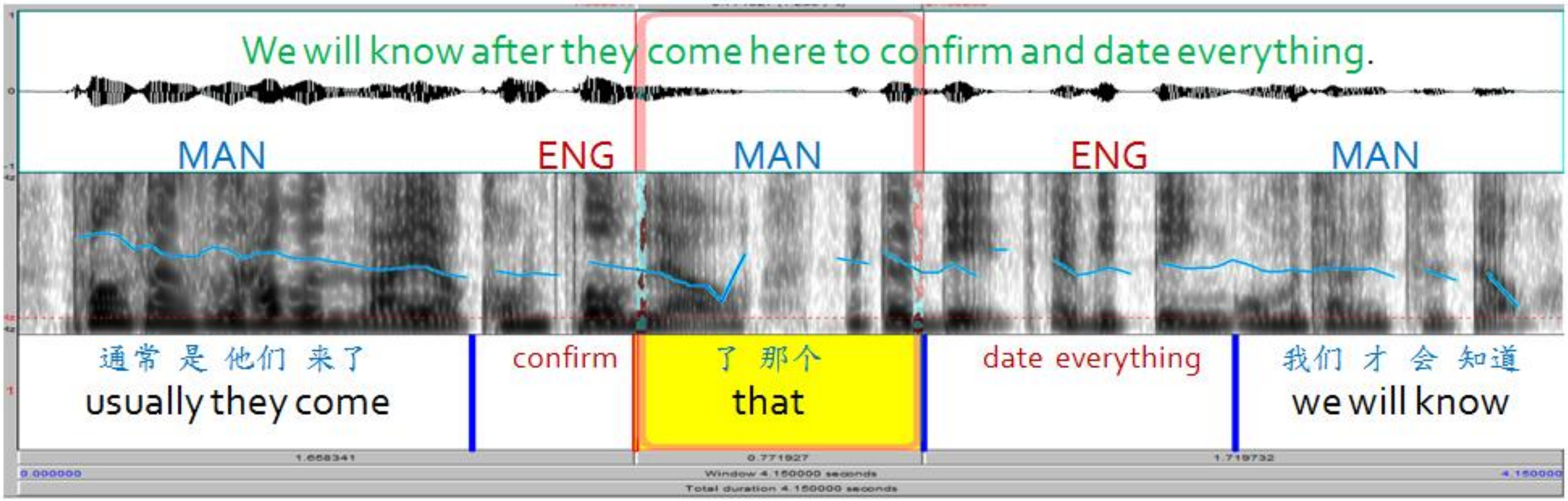


Outline

- Introduction
  - Motivation
  - Seame Corpus
  - Main Contributions
- Factored Language Models
  - Features for Code-Switching Speech
- Experiments
- Conclusion
  - Summary

Motivation

- Code-Switching (CS) = speech with more than one language
- Exists in multilingual communities or among immigrants



Challenges:
multilingual models and CS training data necessary

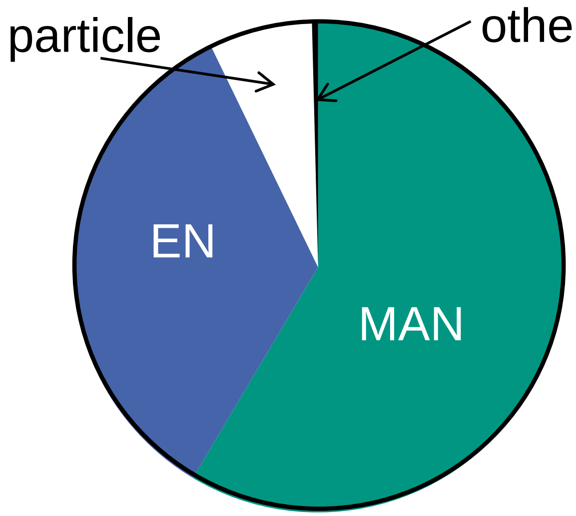
SEAME corpus

- SEAME = South East Asia Mandarin-English
- Conversational speech, recorded from Singaporean and Malaysian speakers by [1]

Challenges

- much CS per utterance (∅: 2.6)
- short monolingual segments (mostly less than 1 sec, 2-4 words)
- not much training data for LM (575k words)

Language distribution



[1] Lyu, D.C. et al.,2010
Originally used: research project 'Code-Switch' (NTU and KIT)

Main contributions

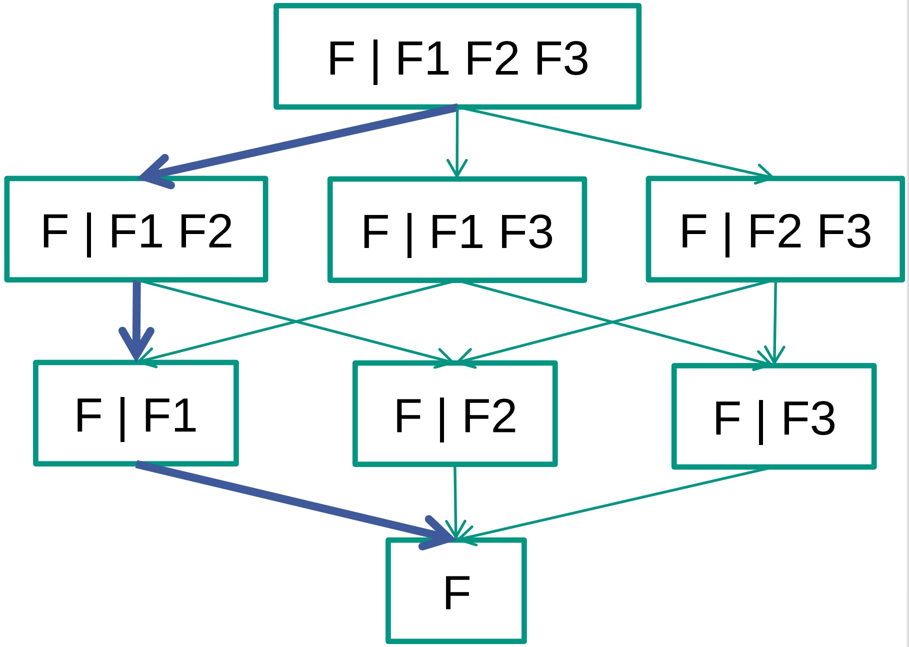
- Investigation of different features for Code-Switching speech
- Integration of factored language models into a dynamic one-pass decoder

Factored Language Models (FLMs) [2]

- Idea: word = feature bundle

$$w_t \equiv \{ f_t^1, f_t^2, \ldots, f_t^K \}$$

- Good e.g. in the case of
  - Rich morphology
  - Few training data
    => applicable to CS task
- Generalized backoff



[2] Bilmes, J. and Kirchhoff, K., 2003

Features: Words, POS, LID

Model	PPL dev	PPL eval
Baseline (3-gram)	268.39	282.86
POS	260.70	267.86
LID	263.24	267.63
POS + LID	**257.62**	**264.20**

- Problems:
  - POS tagging of CS speech: challenging
  - Accuracy of POS tagger: unknown

→ different clustering method may be more robust

Features: Brown Word Clusters

- Clusters based on word distributions in text [3]
  - minimize average mutual information loss
- Best number of classes in terms of PPL: 70

Model	PPL dev	PPL eval
Baseline (3-gram)	268.39	282.86
POS + LID	257.62	264.20
Brown clusters	257.17	265.50
Brown clusters + POS	**249.00**	**255.34**
Brown cl + POS + LID	251.39	259.05

- So far: clusters based on syntax or word distributions
  ➔ next step: semantic features

[3] Brown, P.F. Et al. ,1992

Features: Open Class Words

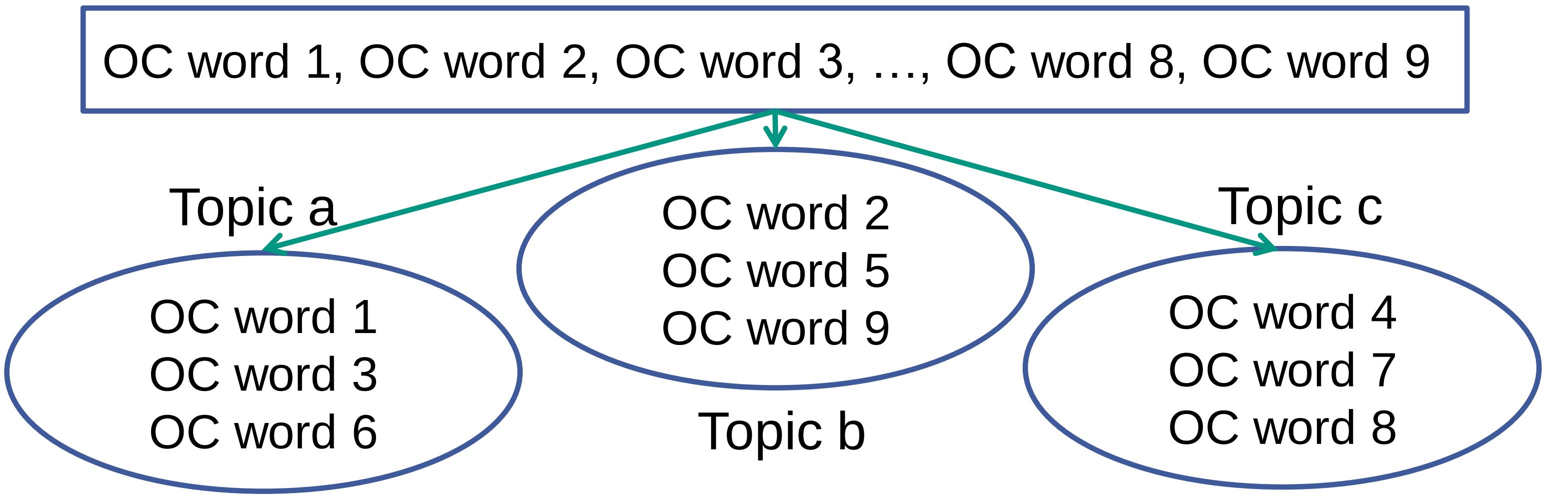
- Definition: content words, e.g. nouns, verbs, adverbs
  - “open” because class can be extended with new words, e.g. “Bollywood”

  ➔ open class words indicate semantic of sentence

Model	PPL dev	PPL eval
Baseline 3-gram	268.39	282.86
Brown clusters + POS	249.00	255.34
Last oc word per speaker + Brown clusters + POS	**247.18**	**252.37**

Features: Open Class Word Clusters

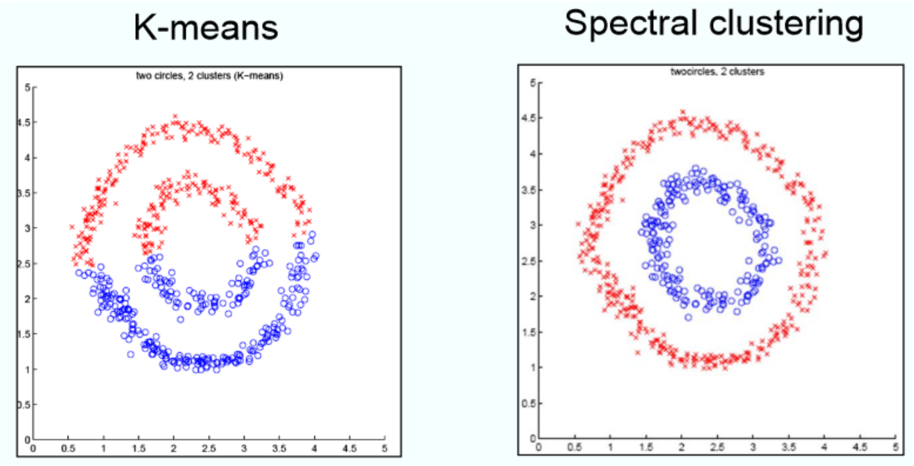
- Idea:



- Semantic clusters in comparison to distribution based clusters (oc Brown clusters)

Features: Semantic OC Word Clusters

- Clustering of open class word vectors
  - RNNLMs learn syntactic and semantic similarities [4]
  - RNNLMs represent words as vectors
  - ➔ apply clustering to these word vectors
    - k-means clustering
    - spectral clustering



[4] Mikolov, T. et al., 2013

Features: Semantic OC Word Clusters

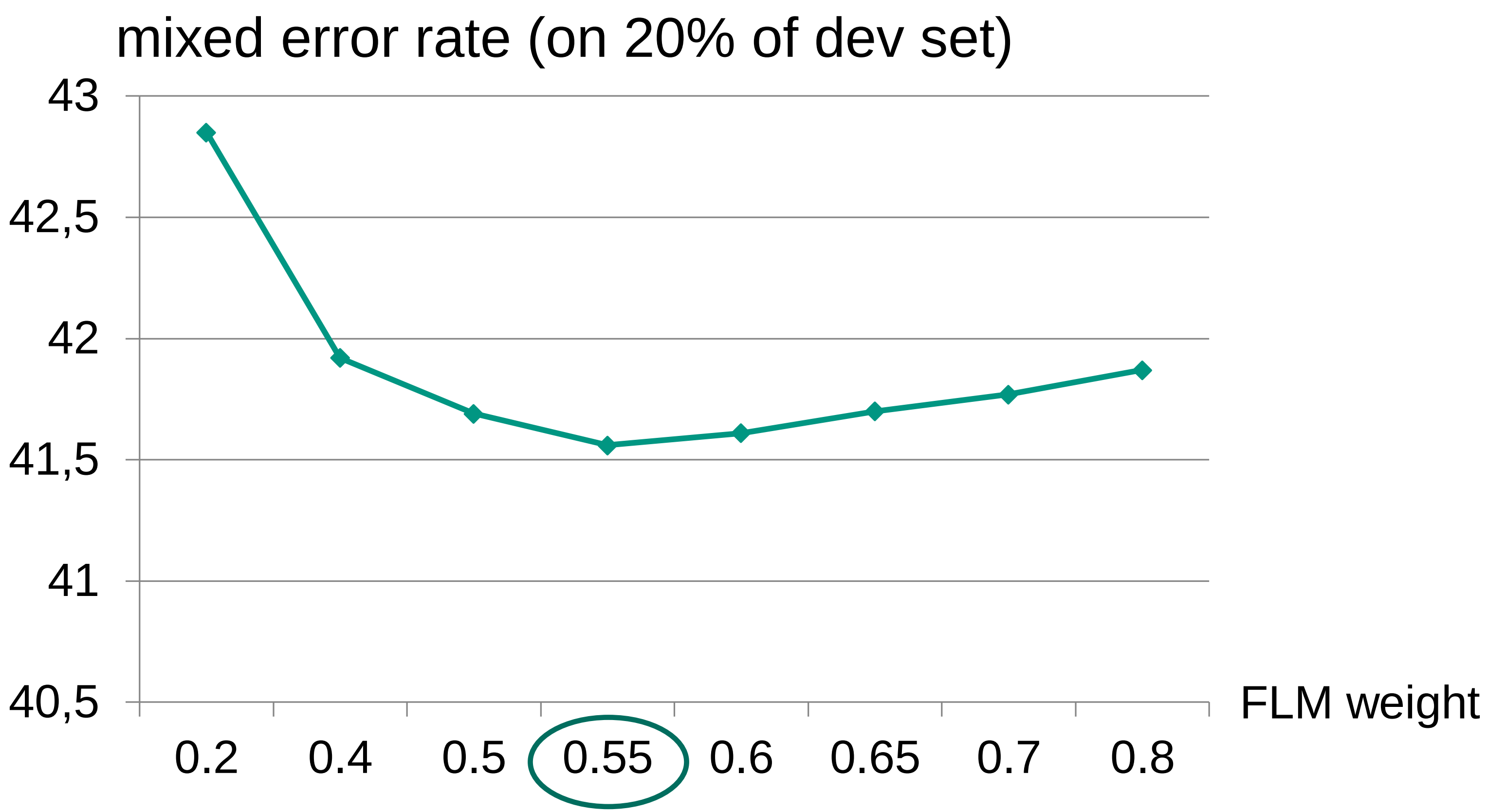
- Experiments with different
  - Clustering methods
    - Brown, k-means, Spectral Clustering
  - Monolingual and bilingual clusters
    - Monolingual Clusters
      - Based on English and Mandarin Gigaword data (2005)
    - Bilingual Clusters
      - Based on CS text
      - Mixed lines of Gigaword data
  - different numbers of clusters

➔ Lowest perplexity (247.24, but unclustered oc words: 247.18):

- Spectral Clustering
- Bilingual Clusters
- 800 OC word clusters

FLMs: Decoding Experiments

- Interpolation weight of FLM and n-gram



FLMs: Decoding Experiments (2)

- Decoding results

Model	MER dev	MER eval
Baseline 3-gram	39.96%	34.31%
POS	39.47%	33.46%
POS + LID	39.66%	33.30%
Brown clusters	39.45%	33.93%
Brown clusters + POS	**39.30%**	33.60%
Brown clusters + POS + LID	39.39%	33.16%
OC words + Brown clusters + POS	39.33%	**33.15%**
OC clusters + Brown clusters + POS	**39.30%**	33.16%

Conclusion

- Summary
  - Best features in terms of FLM perplexity:
    words + POS + Brown clusters + oc words
  - Relative PPL reduction of up to 10.8% (eval)
  - Best features in terms of MER:
    words + POS + Brown clusters (+ oc clusters)
  - Relative MER reduction of up to 3.4% (eval)

THANK YOU FOR YOUR ATTENTION!

References

[1] S. Poplack, *Syntactic structure and social function of code-switching*, vol. 2, Centro de Estudios Puertorriqueños,[City University of New York], 1978.

[2] E.G. Bokamba, "Are there syntactic constraints on code-mixing?," *World Englishes*, vol. 8, no. 3, pp. 277–292, 1989.

[3] P. Muysken, *Bilingual speech: A typology of code-mixing*, vol. 11, Cambridge University Press, 2000.

[4] P. Auer, "From codeswitching via language mixing to fused lects toward a dynamic typology of bilingual speech," *International Journal of Bilingualism*, vol. 3, no. 4, pp. 309–332, 1999.

[5] N.T. Vu, H. Adel, and T. Schultz, "An investigation of code-switching attitude dependent language modeling," in *Proc. of SLSP*, 2013.

[6] S. Poplack, "Sometimes I'll start a sentence in spanish y termino en español: toward a typology of code-switching," *Linguistics*, vol. 18, no. 7-8, pp. 581–618, 1980.

[7] T. Solorio and Y. Liu, "Learning to predict code-switching points," in *Proc. of the Conference on Empirical Methods in Natural Language Processing*. ACL, 2008, pp. 973–981.

[8] J.Y.C. Chan, PC Ching, T. Lee, and H. Cao, "Automatic speech recognition of Cantonese-English code-mixing utterances," in *Proc. of Interspeech*, 2006.

[9] H. Adel, N.T. Vu, F. Kraus, T. Schlippe, H. Li, and T. Schultz, "Recurrent neural network language modeling for code switching conversational speech," in *Proc. of ICASSP*. IEEE, 2013.

[10] K. Duh and K. Kirchhoff, "Automatic learning of language model structure," in *Proc. of the 20th international conference on Computational Linguistics*. ACL, 2004, p. 148.

[11] A. El-Desoky, R. Schlüter, and H. Ney, "A hybrid morphologically decomposed factored language models for arabic LVCSR," in *Proc. of HLT-NAACL*. ACL, 2010, pp. 701–704.

[12] K. Kirchhoff, J.A. Bilmes, and K. Duh, "Factored language models tutorial," Tech. Rep. UWEETR-2008-004, University of Washington, EE Department, 2007.

[13] H. Adel, N.T. Vu, and T. Schultz, "Combination of recurrent neural networks and factored language models for code-switching language modeling," in *Proc. of ACL*, 2013.

References

[14] T. Mikolov, W.-T. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proc. of HLT-NAACL*. ACL, 2013, pp. 746–751.

[15] J.A. Bilmes and K. Kirchhoff, "Factored language models and generalized parallel backoff," in *Proc. of HLT-NAACL*. ACL, 2003, pp. 4–6.

[16] T. Schultz, P. Fung, and C. Burgmer, "Detecting code-switch events based on textual features," Diploma thesis, 2009.

[17] C.M. Scotton, *Duelling languages: Grammatical structure in codeswitching*, Oxford University Press, 1997.

[18] K. Toutanova, D. Klein, C.D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proc. of HLT-NAACL*. ACL, 2003, pp. 173–180.

[19] H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C.D. Manning, "A conditional random field word segmenter," *fourth SIGHAN Workshop on Chinese Language Processing*, 2005.

[20] P.F. Brown, P.V. Desouza, R.L. Mercer, V.J.D. Pietra, and J.C. Lai, "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[21] V. Fromkin, *An introduction to language*, Cengage Learning, 2013.

[22] I.S. Dhillon, Y. Guan, and B. Kulis, "Weighted graph cuts without eigenvectors a multilevel approach," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 11, pp. 1944–1957, 2007.

[23] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[24] D.C. Lyu, T.P. Tan, E.S. Chng, and H. Li, "An analysis of a Mandarin-English code-switching speech corpus: SEAME," *Age*, vol. 21, pp. 25–8, 2010.

[25] N.T. Vu, F. Metze, and T. Schultz, "Multilingual bottleneck features and its application for under-resourced languages," in *Proc. of SLTU*, 2012.

[26] "CMU pronouncation dictionary for English," `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`.

[27] R. Hsiao, M. Fuhs, Y.C. Tam, Q. Jin, and T. Schultz, "The CMU-InterACT 2008 mandarin transcription system," in *Proc. of ICASSP*, 2008.

[28] W. Chen, Y. Tan, E. Chng, and H. Li, "The development of a Singapore English call resource," *Oriental COCOSDA, Nepal*, 2010.

[29] A. Stolcke et al., "SRILM-an extensible language modeling toolkit," in *Proc. of SLP*, 2002, vol. 2, pp. 901–904.

[30] 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 *Proc. of ICASSP*. IEEE, 2012, pp. 4889–4892.