Challenges of Computational Processing of Code-Switching

Özlem Çetinoğlu Sarah Schulz Ngoc Thang Vu
IMS, University of Stuttgart

2nd Workshop on Computational Approaches to Code Switching
1 November 2016

Mixing Languages

(Photo: Craig Morey/Flickr)

- Acting multilingual, that is, mixing languages is commonly observed among multilingual speakers [Auer and Wei 2007]

Mixing Languages

- Extensively studied from social and linguistic perspectives
  [Poplack 1980, Myers-Scotton 1993, Muysken 2000, Auer and Wei 2007, Bullock and Toribio 2012]

- Different use of terminology
  - inter-sentential vs. intra-sentential
  - *code-mixing* for intra-sentential alternations
  - either *code-mixing* or *code-switching* for all types of mixing
  - borrowing vs. code-mixing

- Code-switching for all types of mixing

Computational Processing

Normalisation

Language Modelling

Language Identification

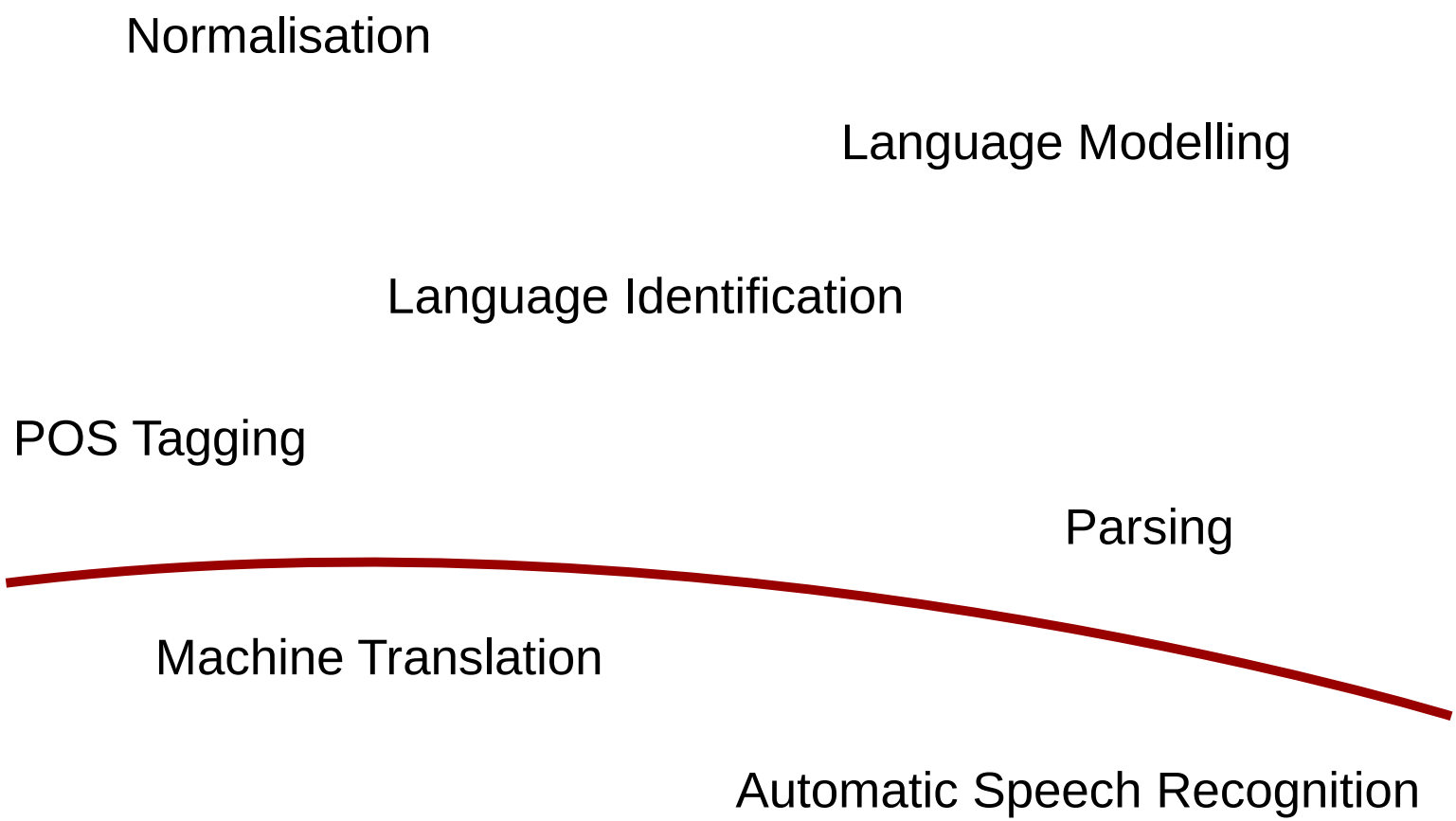
POS Tagging

Parsing

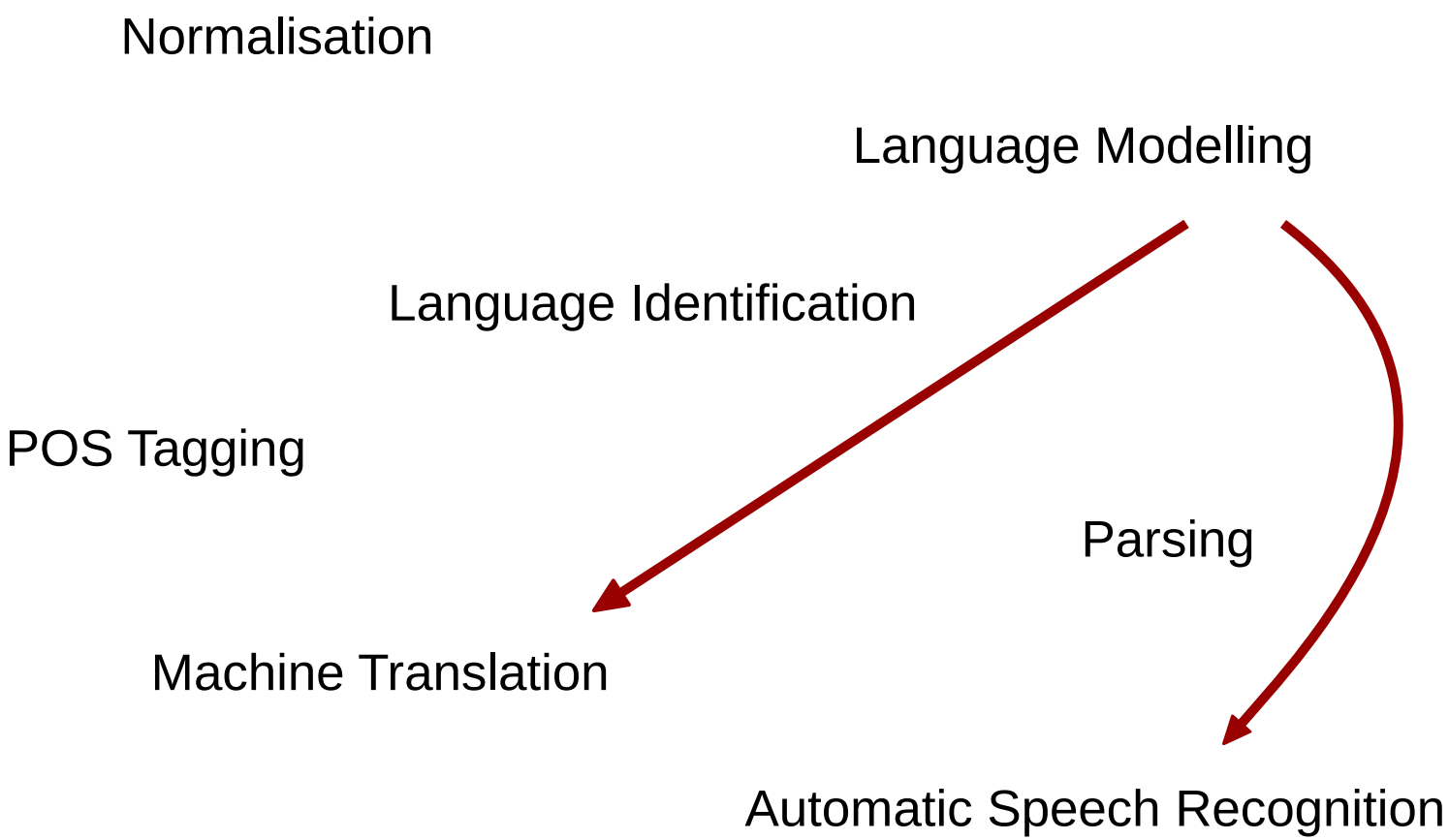
Machine Translation

Automatic Speech Recognition

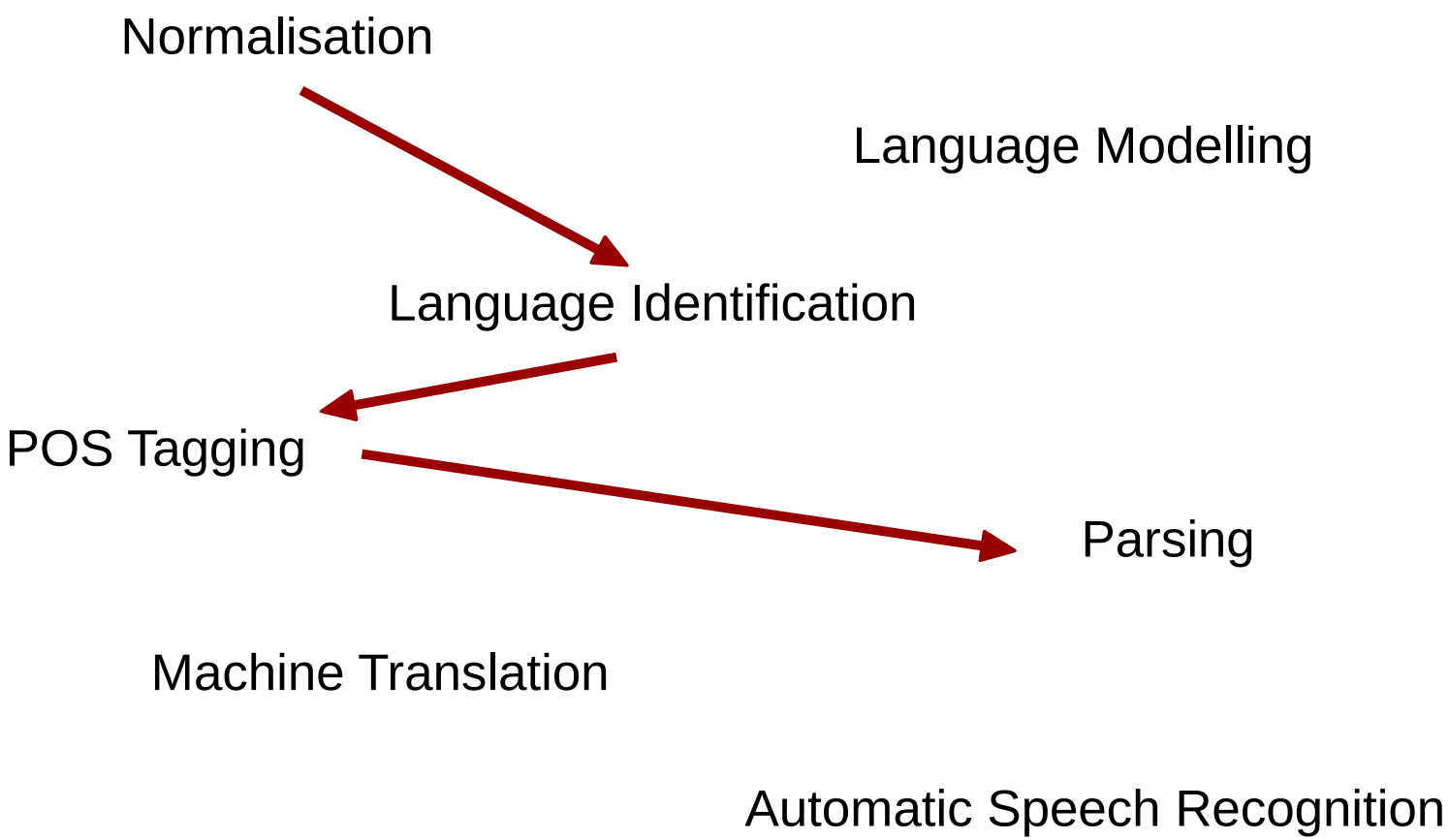
Computational Processing



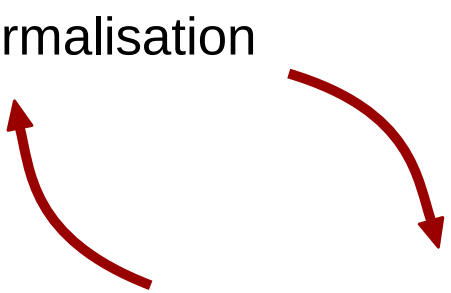
Computational Processing



Computational Processing



Computational Processing



Language Modelling

POS Tagging

Parsing

Machine Translation

Automatic Speech Recognition

Nature of the Data

- Spoken data [Solorio and Liu 2008, Lyu et al. 2015, Yılmaz et al. 2016]
- Historical text [Schulz and Keller 2016]
- Social media [Nguyen and Doğruöz 2013, Barman et al. 2014, Vyas et al. 2014] [Solorio et al. 2014, Choudhury et al. 2014, Jamatia et al. 2015] [Çetinoğlu 2016, Samih and Maier 2016]

(1) *vette spellllllllll* bir girdimmi cikamiyomm
fat game once enter.Past.1Sg leave.Neg.Prog.1Sg
'Awesome game, once I enter I cannot leave.'

Nature of the Data

- Spoken data [Solorio and Liu 2008, Lyu et al. 2015, Yılmaz et al. 2016]
- Historical text [Schulz and Keller 2016]
- Social media [Nguyen and Doğruöz 2013, Barman et al. 2014, Vyas et al. 2014] [Solorio et al. 2014, Choudhury et al. 2014, Jamatia et al. 2015] [Çetinoğlu 2016, Samih and Maier 2016]

(2) *vette* *spelllllllll* bir girdimmi cikamiyomm
fat game once enter.Past.1Sg leave.Neg.Prog.1Sg
'Awesome game, once I enter I cannot leave.'

Accessing the Data

- Spoken Data
  - Prior consent from sources
  - Natural language usage
  - Self-awareness
- Social Media Data
  - Post consent from sources
  - License issues
  - Noisier environment?

! How to collect big sizes of data?

! How to share?

Normalisation

this workshop IS gr8!!!! $\rightarrow$ This workshop is great!

- Standardising text that deviates from some agreed-upon (or canonical) form
  - not a valid word
  - valid word, but a wrong one in context

Normalisation

- Highly relevant to CS due to the nature of the data
- ! Additional challenge: context according to the language

(3) ***meisten*** *kıyımıza* *vurmuş* *olmasi*
Meis.Abl(TR)/mostly(DE) shore.P1pl.Dat hit.EvidPast be.Inf
muhtemel :)
possible

'It is possible that it hit our shore from Meis.'/'Mostly it is possible that it hit our shore.'

- meisten → Meis'ten (from Meis)
  meisten → Meistens (Mostly)

Normalisation

- Neural net-based translation architecture [Zhang et al. 2014]
- Monolingual language models with context depending on neighbouring language [Dutta et al. 2015]

- Text in Roman script, resources in another script
  - Punjabi-English [Kaur and Singh 2015]
  - Hindi-English and Bengali-English [Sarkar 2016]
- Different approaches
  - Use training set as a resource [Barman et al. 2014]
  - Romanise the resources [Das and Gambäck 2014]
  - Transliterate back to the other script [Vyas et al. 2014]

Language Modelling

$$
\begin{gathered}
\ldots \\
P(\text{is} \mid \text{this, workshop}) \\
P(\text{great} \mid \text{workshop, is}) \\
\ldots
\end{gathered}
$$

- Consider mixed text as an individual language, use existing methods
  - Tokenisation and normalisation as preprocessing
- Inter-sentential CS:
  - Monolingual models from CS language pair
  - Model interpolation

Language Modelling

- Intra-sentential CS: only CS data
  - ! When do people code-switch?
  - ! How to incorporate this information?

- Language ID and POS information [Adel et al. 2013a, 2015, 2013b]
  - Mandarin and English POS tags that trigger a code-switching point [Adel et al. 2013a]
  - Man $\rightarrow$ En: DT 40.44%
  - En $\rightarrow$ Man: NN 49.07% , NNS 40.82%
- Functional head constraints [Li and Fung 2012, 2014]

- ! Intra-word CS: Additional out-of-vocabulary problem

Language Identification

This workshop is great → EN
Dieser Workshop ist großartig → DE

- Highly accurate on monolingual data
  - Up to 100%
  - Harder to discriminate similar languages [Zampieri et al. 2014]
- ! Mixed text: More than one languages in input, one ID is not enough
  - 💡 Token level identification

Language Identification

- Most well-studied task among computational CS approaches:
  - Relatively more annotated data
  - A preprocessing step for more complex tasks
  - Shared tasks [Solorio et al. 2014, Choudhury et al. 2014]

- High accuracy language identifiers has lower performance on CS data [Nguyen and Doğruöz 2013, Volk and Clematide 2014]

- Approaches tailored to mixed data
  - Utilising monolingual dictionaries
  - Machine learning on annotated CS data [Lignos and Marcus 2013, Nguyen and Doğruöz 2013, Voss et al. 2014, Das and Gambäck 2014, Barman et al. 2014, Solorio et al. 2014]
  - Accuracies in mid-90s; lower F1 (80-85%) for some language pairs

Language Identification

- Closely related languages – linguistically or historically
  - Modern Standard Arabic and dialects [Elfardy and Diab 2012, Samih and Maier 2016]
  - Frisian-Dutch [Yılmaz et al. 2016]
  - English-Hindi, English-Bengali [Das and Gambäck 2014, Vyas et al. 2014, Barman et al. 2014]
- Hard to find a clear distinction between CS and borrowing
  - ! Resources and annotation guidelines might conflict
  - ! Annotators agreement is low
  - ! Also observed when a third language has influence on one or both of CS languages

- ! Intra-word CS: How to annotate?
  - 💡 No special tag (no or very infrequent occurence)
  - 💡 `Mixed` tag [Maharjan et al. 2015, Barman et al. 2014, Çetinoğlu 2016]
  - 💡 Fine-grained tag set [Das and Gambäck 2014]

POS Tagging

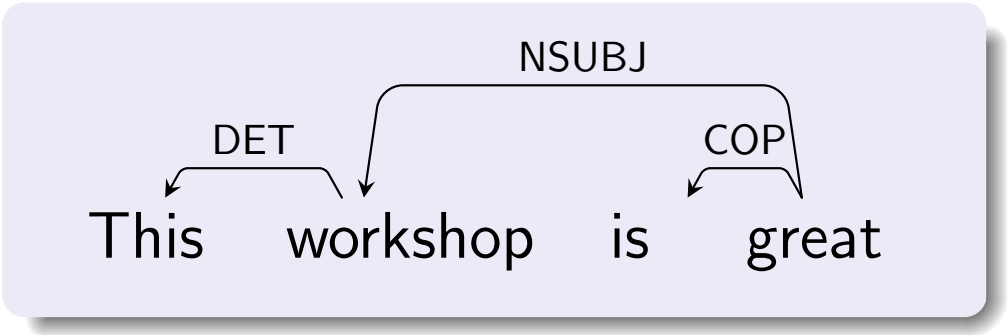
This	workshop	is	great
DET	NOUN	VERB	ADJ

- State-of-the-art on monolingual data: 97% accuracy
- POS-annotated CS corpora
  - Languages: En-Es, En-Hi(x4), En-Bn, En-Ta, De-Tr, midEn-La
  - Tokens: 3k-38k
  - Tag Sets: 1 language-specific, others: universal tag sets
    [Solorio and Liu 2008, Vyas et al. 2014, Jamatia et al. 2015, Çetinoğlu and Çöltekin 2016, Sharma et al. 2016, Schulz and Keller 2016]

POS Tagging

- Monolingual off-the-shelf taggers are not suitable
  - En tagger 54% Es tagger 26% accuracy on En-Es [Solorio and Liu 2008]
- Approaches tailored to mixed data
  - Choosing between monolingual tagger outputs based on probabilities
  - Utilising monolingual dictionaries and language models
  - Machine learning on annotated CS data
    [Solorio and Liu 2008, Vyas et al. 2014, Jamatia et al. 2015, Sharma et al. 2016, Schulz and Keller 2016]
- Language ID as useful feature
  - Language identification as a preprocessing step
- Accuracies in the 65-75% range
  - Exception: 93.48% on En-Es [Solorio and Liu 2008], partly due to 62.5% monolingual English sentences

Parsing



- Advanced substantially over the last decade
  - e.g. English dependency parsing above 93% UAS [Kiperwasser and Goldberg 2016]

Parsing

- Parsing CS text
  - Theoretical framework [Joshi 1982]
  - Rule-based HPSG prototype [Goyal et al. 2003]
  - But no statistical parsers $\leftarrow$ no CS treebanks
- Chunking for Hindi-English social media [Sharma et al. 2016]
- Parsing 10 English-Spanish tweets [Vilares et al. 2016]
  - English and Spanish data are combined to train the POS tagger and parser

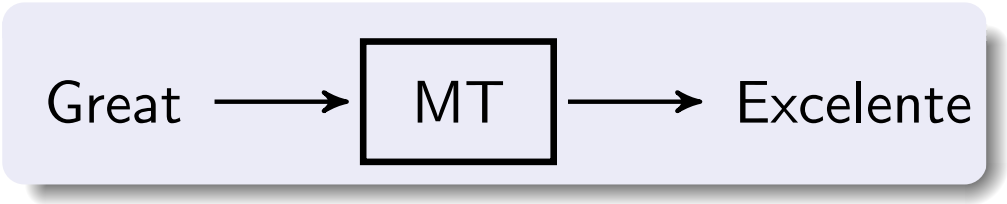
Parsing

- ! Error propogation from previous steps
- ! Syntactic constructions that are not native to monolingual languages

(4) birkaç *Aufgabe*ler yaptık arkadaşla
a few assignment(DE).Pl(TR) make.Past.1Pl friend.Sg.Ins
'We made a few assignments with a friend.'

- *birkaç Aufgabeler* does not follow Turkish syntax
- Turkish syntax: singular noun in NP, German syntax: plural

Machine Translation



- Parallel text for phrase tables and translation probabilities, monolingual data for language models
  - ! Mixed text: high number of unknowns, low probabilities for translations

Machine Translation

- Foreign word translation into the source language before translation into the target language [Sinha and Thakur 2005]
  - ! Foreign word detection, language identification, normalisation

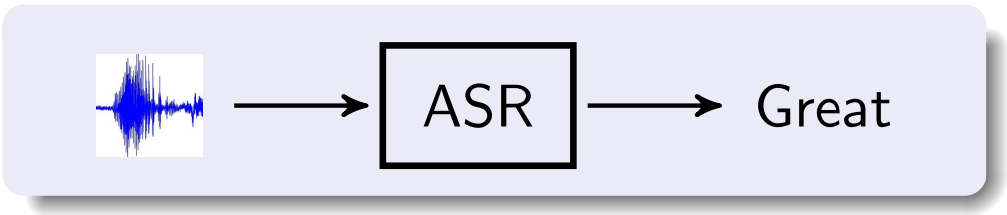
- Intra-word CS: Morphological analysis to separate the stem from suffixes, then a dictionary lookup [Manandise and Gdaniec 2011]

(5)

anticooldad →	anticool: dad	
	cooldad : anti	
	cool: anti, dad	→ anti-coolness

- ! What about syntax of the monolingualised source language?

Automatic Speech Recognition

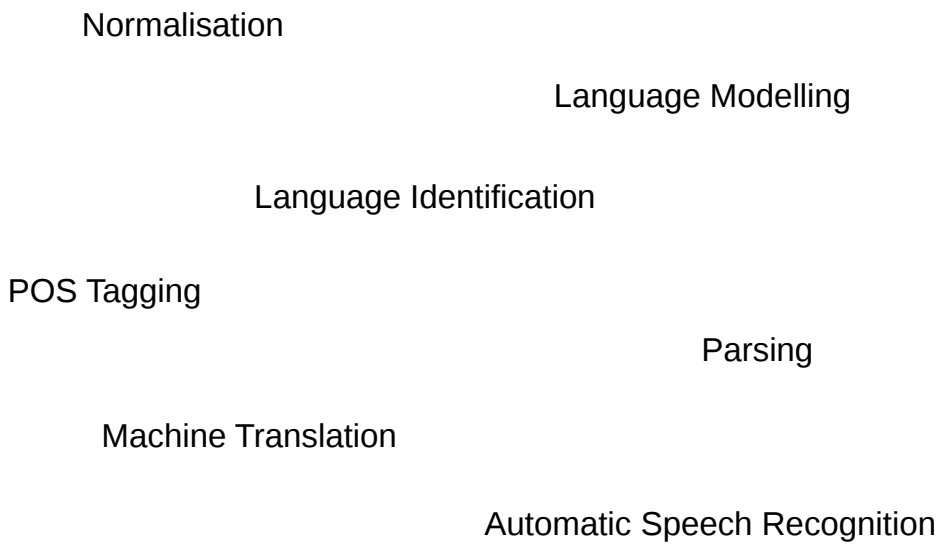


- Three major components
  - A pronunciation dictionary, a language model and an acoustic model
- Dealing with CS speech [Vu et al. 2012]
- Split input into monolingual parts, apply monolingual recognisers
  - Language identification mistakes if segments are short (e.g. < 3s)
  - Lost context information
  - Intra-word CS
- Multilingual components

Automatic Speech Recognition

- Pronunciation dictionary
  - Collection of words and phoneme sequences which describe how a word is pronounced
  - ! CS changes pronunciation due to articulation effect
  - ! Pronunciation of intra-word CS
- Acoustic modelling
  - Estimating the probability of a sound state given a speech frame
  - ! Phonetic transfer phenomenon
- Merging phoneme sets for bilingual sound state models [Vu et al. 2012]
- Creating phone clusters and linear interpolation of their sound state models [Li and Fung 2013]
- Integrating language identification into ASR in the speech frame level [Weiner et al. 2012]

Overview



- Need for more data
  - New lexical and syntactic structures
- Few resources
  - Language ID, POS tagging, ASR
- Artificial data
- Linguistic insights
- Joint approaches

Thanks!

Questions?

References I

Adel, H., Vu, N. T., Kirchhoff, K., Telaar, D., and Schultz, T. (2015). Syntactic and semantic features for code-switching factored language models. *IEEE/ACM TASL*, 23(3).

Adel, H., Vu, N. T., Kraus, F., Schlippe, T., Schultz, T., and Li, H. (2013a). Recurrent neural network language modeling for code switching conversational speech. In *Proceedings of ICASSP*.

Adel, H., Vu, N. T., and Schultz, T. (2013b). Combination of recurrent neural networks and factored language models for code-switching language modeling. In *Proceedings of ACL*.

Auer, P. and Wei, L. (2007). *Handbook of multilingualism and multilingual communication*, volume 5. Walter de Gruyter.

Barman, U., Das, A., Wagner, J., and Foster, J. (2014). Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the CodeSwitch Workshop*.

Bullock, B. E. and Toribio, A. J. (2012). *The Cambridge handbook of linguistic code-switching*. Cambridge University Press.

Çetinoğlu, O. (2016). A Turkish-German code-switching corpus. In *Proceedings of LREC*.

Çetinoğlu, Ö. and Çöltekin, Ç. (2016). Part of speech tagging of a turkish-german code-switching corpus. In *Proceedings of LAW-X*.

Choudhury, M., Chittaranjan, G., Gupta, P., and Das, A. (2014). Overview of fire 2014 track on transliterated search. In *Proceedings of FIRE*.

References II

Das, A. and Gambäck, B. (2014). Code-mixing in social media text: the last language identification frontier. *Traitement Automatique des Langues (TAL): Special Issue on Social Networks and NLP*, 54(3).

Dutta, S., Saha, T., Banerjee, S., and Naskar, S. K. (2015). Text normalization in code-mixed social media text. In *Recent Trends in Information Systems (ReTIS)*.

Elfardy, H. and Diab, M. (2012). Simplified guidelines for the creation of large scale dialectal arabic annotations. In *Proceedings of LREC*.

Goyal, P., Mital, M. R., Mukerjee, A., Raina, A. M., Sharma, D., Shukla, P., and Vikram, K. (2003). A bilingual parser for hindi, english and code-switching structures. In *Proceedings of EACL*.

Jamatia, A., Gambäck, B., and Das, A. (2015). Pos tagging for code-mixed english-hindi twitter and facebook chat messages. In *Proc. of RANLP*.

Joshi, A. K. (1982). Processing of sentences with intra-sentential code-switching. In *Proc. of COLING*.

Kaur, J. and Singh, J. (2015). Toward normalizing romanized gurumukhi text from social media. *Indian Journal of Science and Technology*, 8(27).

Kiperwasser, E. and Goldberg, Y. (2016). Simple and accurate dependency parsing using bidirectional LSTM feature representations. *TACL*, 4.

Li, Y. and Fung, P. (2012). Code-switch language model with inversion constraints for mixed language speech recognition. In *Proceedings of COLING*.

References III

Li, Y. and Fung, P. (2013). Improved mixed language speech recognition using asymmetric acoustic model and language model with code-switch inversion constraints. In *Proceedings of ICASSP*.

Li, Y. and Fung, P. (2014). Language modeling with functional head constraint for code switching speech recognition. In *Proceedings of EMNLP*.

Lignos, C. and Marcus, M. (2013). Toward web-scale analysis of the codeswitching. In *Proceedings of of LSA*.

Lyu, D.-C., Tan, T.-P., Chng, E.-S., and Li, H. (2015). Mandarin–english code-switching speech corpus in south-east asia: Seame. *LRE*, 49(3):581–600.

Maharjan, S., Blair, E., Bethard, S., and Solorio, T. (2015). Developing language-tagged corpora for code-switching tweets. In *Proceedings of LAW-9*.

Manandise, E. and Gdaniec, C. (2011). Morphology to the rescue redux: Resolving borrowings and code-mixing in machine translation. In *Proceedings of SFCM*.

Muysken, P. (2000). *Bilingual Speech: A Typology of Code-Mixing*.

Myers-Scotton, C. (1993). *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.

Nguyen, D. and Doğruöz, A. S. (2013). Word level language identification in online multilingual communication. In *Proceedings of EMNLP*.

References IV

Poplack, S. (1980). Sometimes I'll start a sentence in Spanish y termino en Espanol: toward a typology of code-switching. *Linguistics*, 18(7-8).

Samih, Y. and Maier, W. (2016). An arabic-moroccan darija code-switched corpus. In *Proceedings of LREC*.

Sarkar, K. (2016). Part-of-speech tagging for code-mixed indian social media text at ICON 2015. *CoRR*, abs/1601.01195.

Schulz, S. and Keller, M. (2016). Code-switching ubique est - language identification and part-of-speech tagging for historical mixed text. In *Proc. of LaTeCH*.

Sharma, A., Gupta, S., Motlani, R., Bansal, P., Shrivastava, M., Mamidi, R., and Sharma, D. M. (2016). Shallow parsing pipeline - hindi-english code-mixed social media text. In *Proceedings of NAACL*.

Sinha, R. and Thakur, A. (2005). Machine translation of bi-lingual hindi-english (hinglish) text. In *Proceedings of the MT Summit*.

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Ghoneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., and Fung, P. (2014). Overview for the first shared task on language identification in code-switched data. In *Proceedings of the CodeSwitch Workshop*.

Solorio, T. and Liu, Y. (2008). Part-of-Speech tagging for English-Spanish code-switched text. In *Proceedings of EMNLP*.

References V

Vilares, D., Alonso, M. A., and Gómez-Rodríguez, C. (2016). One model, two languages: training bilingual parsers with harmonized treebanks. In *Proc. of ACL*.

Volk, M. and Clematide, S. (2014). Detecting code-switching in a multilingual alpine heritage corpus. In *Proceedings of the CodeSwitch Workshop*.

Voss, C., Tratz, S., Laoudi, J., and Briesch, D. (2014). Finding romanized arabic dialect in code-mixed tweets. In *Proceedings of LREC*.

Vu, N. T., Lyu, D., Weiner, J., Telaar, D., Schlippe, T., Blaicher, F., Chng, E.-S., Schultz, T., and Li, H. (2012). A first speech recognition system for mandarin-english code-switch conversational speech. In *Proc. of ICASSP*.

Vyas, Y., Gella, S., Sharma, J., Bali, K., and Choudhury, M. (2014). Pos tagging of english-hindi code-mixed social media content. In *Proceedings of EMNLP*.

Weiner, J., Vu, N. T., Telaar, D., Metze, F., Schultz, T., Lyu, D., Chng, E.-S., and Li, H. (2012). Integration of language identification into a recognition system for spoken conversations containing code-switches. In *Proceedings of SLTU*.

Yılmaz, E., Andringa, M., Kingma, S., Dijkstra, J., der Kuip, F. V., de Velde, H. V., Kampstra, F., Algra, J., van den Heuvel, H., and van Leeuwen, D. (2016). A longitudinal bilingual frisian-dutch radio broadcast database designed for code-switching research. In *Proceedings of LREC*.

References VI

Zampieri, M., Tan, L., Ljubešić, N., and Tiedemann, J. (2014). A report on the dsl shared task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, Dublin, Ireland.

Zhang, Q., Chen, H., and Huang, X. (2014). Chinese-english mixed text normalization. In *Proc. of WSDM*.