From Language to Family and Back

Native Language and Language Family Identification from English Text

Ariel Stolerman Aylin Caliskan Islam Rachel Greenstadt

Dept. of Computer Science
Drexel University
Philadelphia, PA

NAACL HLT SRW 2013

Outline

Outline

Outline

Outline

Problem Statement and Motivation

- Stylometry: the study of linguistic style
  Question: *Who wrote this document?*
- Utilize stylometric analysis to characterize anonymous authors
  - Age, Gender, Native language...
  - Profiling criminals / terrorists / unveil identities

Problem Statement and Motivation

- Stylometry: the study of linguistic style
  Question: *Who wrote this document?*
- Utilize stylometric analysis to characterize anonymous authors
  - Age, Gender, Native language...
  - Profiling criminals / terrorists / unveil identities

Problem Statement and Motivation

- Stylometry: the study of linguistic style
  Question: *Who wrote this document?*
- Utilize stylometric analysis to characterize anonymous authors
  - Age, Gender, Native language...
  - Profiling criminals / terrorists / unveil identities

Problem Statement and Motivation

- Stylometry: the study of linguistic style
  Question: *Who wrote this document?*
- Utilize stylometric analysis to characterize anonymous authors
  - Age, Gender, Native language...
  - Profiling criminals / terrorists / unveil identities

Problem Statement and Motivation

- Definitions
  - L1: native language
  - L2: non-native language
  - LF: language family of L1
- Question of interest:
  Given English text by an unknown author, what is his/her native language(s)?
  - L1-L2 transfer effect → LF-L2 transfer effect?
  - Increase L1-ID via LF-ID?
    - Yes

Problem Statement and Motivation

- Definitions
  - L1: native language
  - L2: non-native language
  - LF: language family of L1
- Question of interest:
  Given English text by an unknown author, what is his/her native language(s)?
  - L1-L2 transfer effect → LF-L2 transfer effect?
  - Increase L1-ID via LF-ID?
    - Yes

Problem Statement and Motivation

- Definitions
  - L1: native language
  - L2: non-native language
  - LF: language family of L1
- Question of interest:
  Given English text by an unknown author, what is his/her native language(s)?
  - L1-L2 transfer effect $\rightarrow$ LF-L2 transfer effect?
  - Increase L1-ID via LF-ID?
    - Yes

Problem Statement and Motivation

- Definitions
  - L1: native language
  - L2: non-native language
  - LF: language family of L1
- Question of interest:
  Given English text by an unknown author, what is his/her native language(s)?
  - L1-L2 transfer effect $\rightarrow$ LF-L2 transfer effect?
  - Increase L1-ID via LF-ID?
    - Yes

Outline

Related Work

- Choueiter et al.: L1 and accent identification from speech [CZN08]
- L1 identification from text
  - Koppel et al. [KSZ05a, KSZ05b]
  - [EGP+07, vH08, WDJ11, BH12]...
- None consider LF effects and utilization for L1-ID

Related Work

- Choueiter et al.: L1 and accent identification from speech [CZN08]
- L1 identification from text
  - Koppel et al. [KSZ05a, KSZ05b]
  - [EGP+07, vH08, WDJ11, BH12]...
- None consider LF effects and utilization for L1-ID

Outline

Corpus, Features, Classifier

- Corpus: ICLEv2
  - English texts by international learners of English, 16 mother-tongues
  - Used 11 L1 of 3 LF: Romance, Slavic & Germanic
- Feature Sets
  - Basic: Adopted from Koppel et al. [KSZ05a, KSZ05b]
    - Function words, frequent letter bigrams, rare part-of-speech bigrams, common misspellings
  - Extended: added grammar features
    - Frequent POS bigrams
  - Grammar: only grammar features
  - InfoGain: Feature space reduction on Extended
- Measured TPR with SVM classifier + cross-validation

Corpus, Features, Classifier

- Corpus: ICLEv2
  - English texts by international learners of English, 16 mother-tongues
  - Used 11 L1 of 3 LF: Romance, Slavic & Germanic
- Feature Sets
  - Basic: Adopted from Koppel et al. [KSZ05a, KSZ05b]
    - Function words, frequent letter bigrams, rare part-of-speech bigrams, common misspellings
  - Extended: added grammar features
    - Frequent POS bigrams
  - Grammar: only grammar features
  - InfoGain: Feature space reduction on Extended
- Measured TPR with SVM classifier + cross-validation

Corpus, Features, Classifier

- Corpus: ICLEv2
  - English texts by international learners of English, 16 mother-tongues
  - Used 11 L1 of 3 LF: Romance, Slavic & Germanic
- Feature Sets
  - Basic: Adopted from Koppel et al. [KSZ05a, KSZ05b]
    - Function words, frequent letter bigrams, rare part-of-speech bigrams, common misspellings
  - Extended: added grammar features
    - Frequent POS bigrams
  - Grammar: only grammar features
  - InfoGain: Feature space reduction on Extended
- Measured TPR with SVM classifier + cross-validation

Corpus, Features, Classifier

- Corpus: ICLEv2
  - English texts by international learners of English, 16 mother-tongues
  - Used 11 L1 of 3 LF: Romance, Slavic & Germanic
- Feature Sets
  - Basic: Adopted from Koppel et al. [KSZ05a, KSZ05b]
    - Function words, frequent letter bigrams, rare part-of-speech bigrams, common misspellings
  - Extended: added grammar features
    - Frequent POS bigrams
  - Grammar: only grammar features
  - InfoGain: Feature space reduction on Extended
- Measured TPR with SVM classifier + cross-validation

Corpus, Features, Classifier

- Corpus: ICLEv2
  - English texts by international learners of English, 16 mother-tongues
  - Used 11 L1 of 3 LF: Romance, Slavic & Germanic
- Feature Sets
  - Basic: Adopted from Koppel et al. [KSZ05a, KSZ05b]
    - Function words, frequent letter bigrams, rare part-of-speech bigrams, common misspellings
  - Extended: added grammar features
    - Frequent POS bigrams
  - Grammar: only grammar features
  - InfoGain: Feature space reduction on Extended
- Measured TPR with SVM classifier + cross-validation

Corpus, Features, Classifier

- Corpus: ICLEv2
  - English texts by international learners of English, 16 mother-tongues
  - Used 11 L1 of 3 LF: Romance, Slavic & Germanic
- Feature Sets
  - Basic: Adopted from Koppel et al. [KSZ05a, KSZ05b]
    - Function words, frequent letter bigrams, rare part-of-speech bigrams, common misspellings
  - Extended: added grammar features
    - Frequent POS bigrams
  - Grammar: only grammar features
  - InfoGain: Feature space reduction on Extended
- Measured TPR with SVM classifier + cross-validation

Outline

9-Class L1 vs. 3-Class LF

- Questions
  - Best LF-ID method?
  - L1-ID vs. LF-ID performance on same data?
- LF-ID methods
  - Trivial
    - Apply L1-ID
    - Take LF of attributed L1
  - Standalone
    - Apply LF-ID directly (using LF as classes)
  - Combined
    - Set confidence threshold t
    - Apply L1-ID, measure chosen L1 probability p
    - If $p \geq t$: take LF by Trivial
    - If $p < t$: take LF by Standalone
    - Hypothesis: Good L1-ID $\rightarrow$ Good LF-ID from that L1

9-Class L1 vs. 3-Class LF

- Questions
  - Best LF-ID method?
  - L1-ID vs. LF-ID performance on same data?
- LF-ID methods
  - Trivial
    - Apply L1-ID
    - Take LF of attributed L1
  - Standalone
    - Apply LF-ID directly (using LF as classes)
  - Combined
    - Set confidence threshold t
    - Apply L1-ID, measure chosen L1 probability p
    - If $p \geq t$: take LF by Trivial
    - If $p < t$: take LF by Standalone
    - Hypothesis: Good L1-ID $\rightarrow$ Good LF-ID from that L1

9-Class L1 vs. 3-Class LF

- Questions
  - Best LF-ID method?
  - L1-ID vs. LF-ID performance on same data?
- LF-ID methods
  - Trivial
    - Apply L1-ID
    - Take LF of attributed L1
  - Standalone
    - Apply LF-ID directly (using LF as classes)
  - Combined
    - Set confidence threshold t
    - Apply L1-ID, measure chosen L1 probability p
    - If $p \geq t$: take LF by Trivial
    - If $p < t$: take LF by Standalone
    - Hypothesis: Good L1-ID $\rightarrow$ Good LF-ID from that L1

9-Class L1 vs. 3-Class LF

- Questions
  - Best LF-ID method?
  - L1-ID vs. LF-ID performance on same data?
- LF-ID methods
  - Trivial
    - Apply L1-ID
    - Take LF of attributed L1
  - Standalone
    - Apply LF-ID directly (using LF as classes)
  - Combined
    - Set confidence threshold t
    - Apply L1-ID, measure chosen L1 probability p
    - If $p \geq t$: take LF by Trivial
    - If $p < t$: take LF by Standalone
    - Hypothesis: Good L1-ID $\rightarrow$ Good LF-ID from that L1

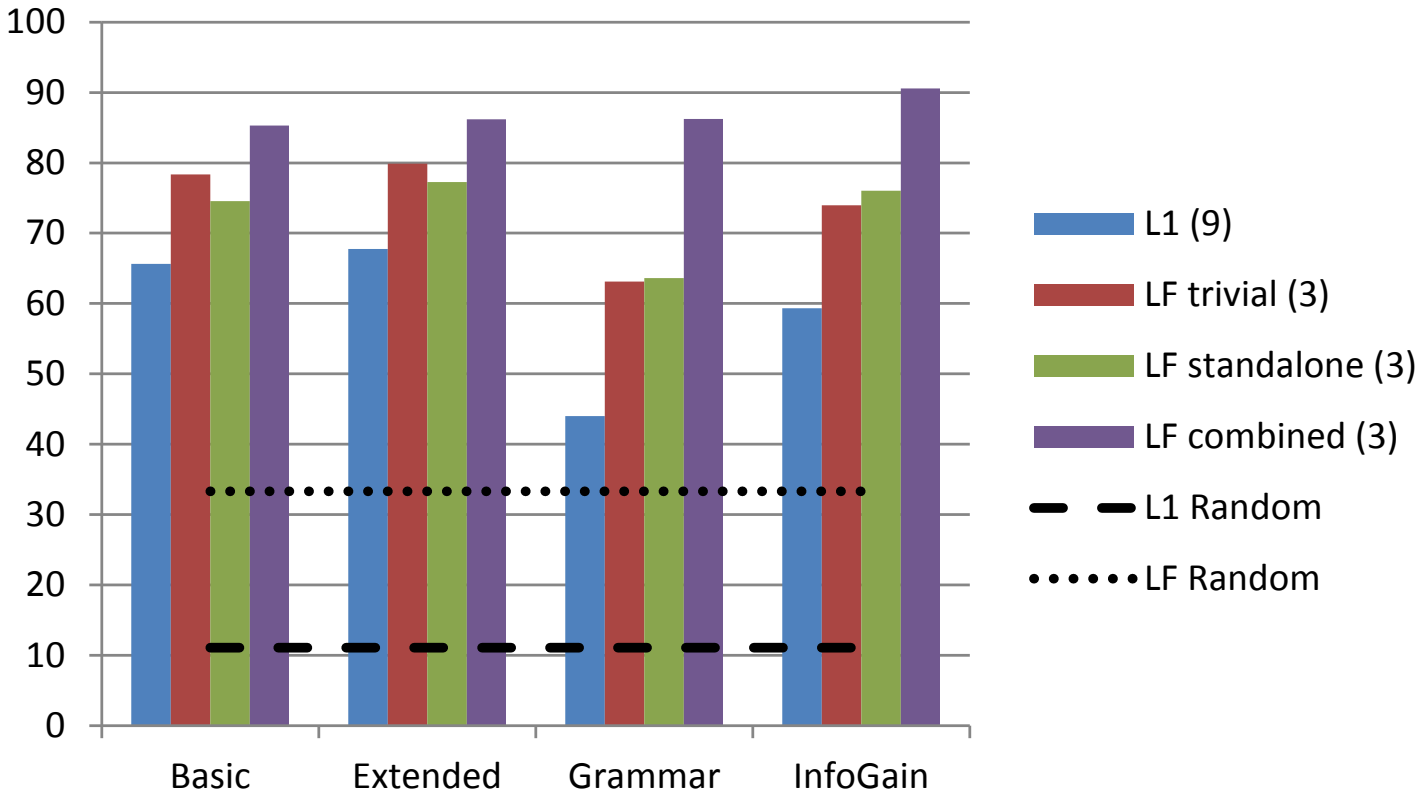
9-Class L1 vs. 3-Class LF

- Questions
  - Best LF-ID method?
  - L1-ID vs. LF-ID performance on same data?
- LF-ID methods
  - Trivial
    - Apply L1-ID
    - Take LF of attributed L1
  - Standalone
    - Apply LF-ID directly (using LF as classes)
  - Combined
    - Set confidence threshold t
    - Apply L1-ID, measure chosen L1 probability p
    - If $p \geq t$: take LF by Trivial
    - If $p < t$: take LF by Standalone
    - Hypothesis: Good L1-ID $\rightarrow$ Good LF-ID from that L1

9-Class L1 vs. 3-Class LF

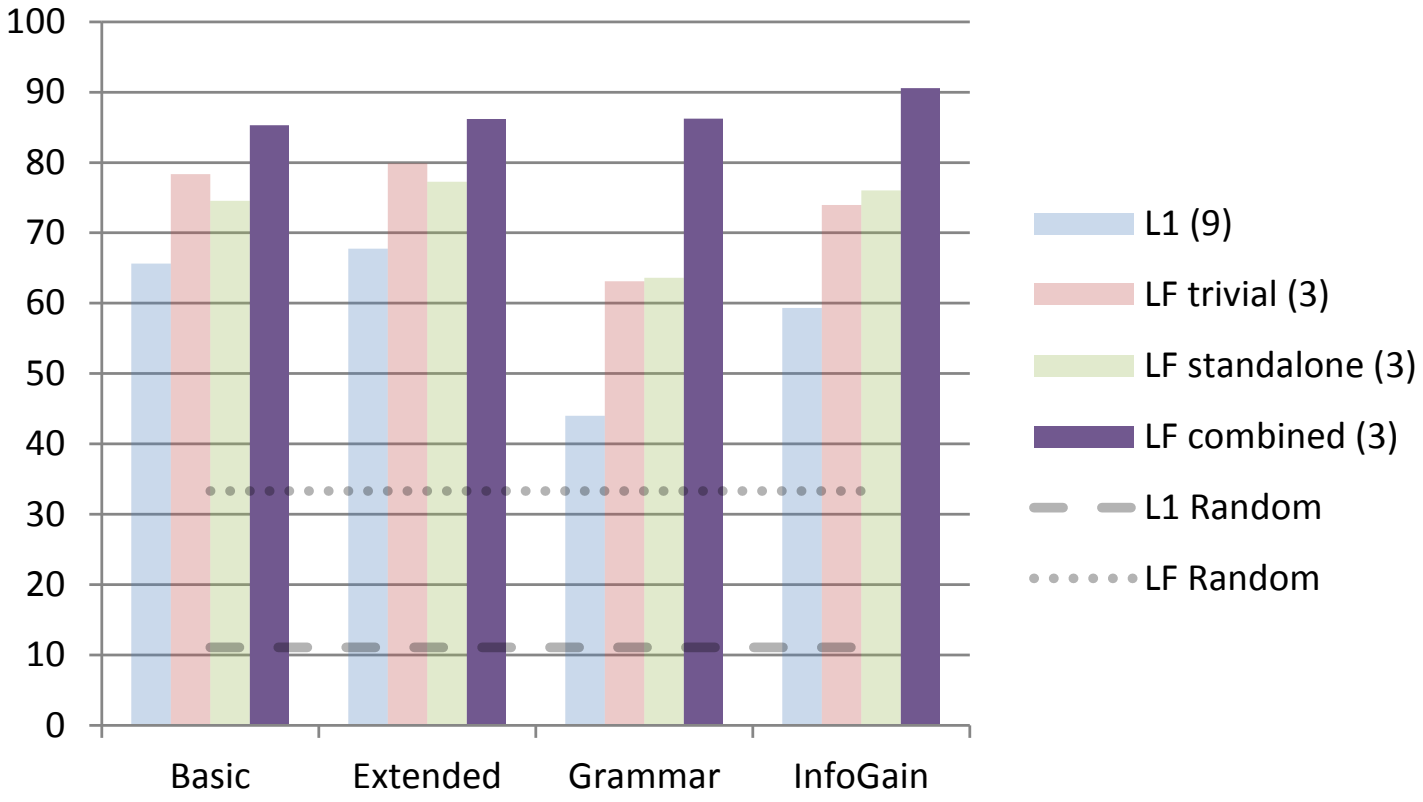
- Questions
  - Best LF-ID method?
  - L1-ID vs. LF-ID performance on same data?
- LF-ID methods
  - Trivial
    - Apply L1-ID
    - Take LF of attributed L1
  - Standalone
    - Apply LF-ID directly (using LF as classes)
  - Combined
    - Set confidence threshold t
    - Apply L1-ID, measure chosen L1 probability p
    - If $p \geq t$: take LF by Trivial
    - If $p < t$: take LF by Standalone
    - Hypothesis: Good L1-ID $\rightarrow$ Good LF-ID from that L1

9-Class L1 vs. 3-Class LF – Results

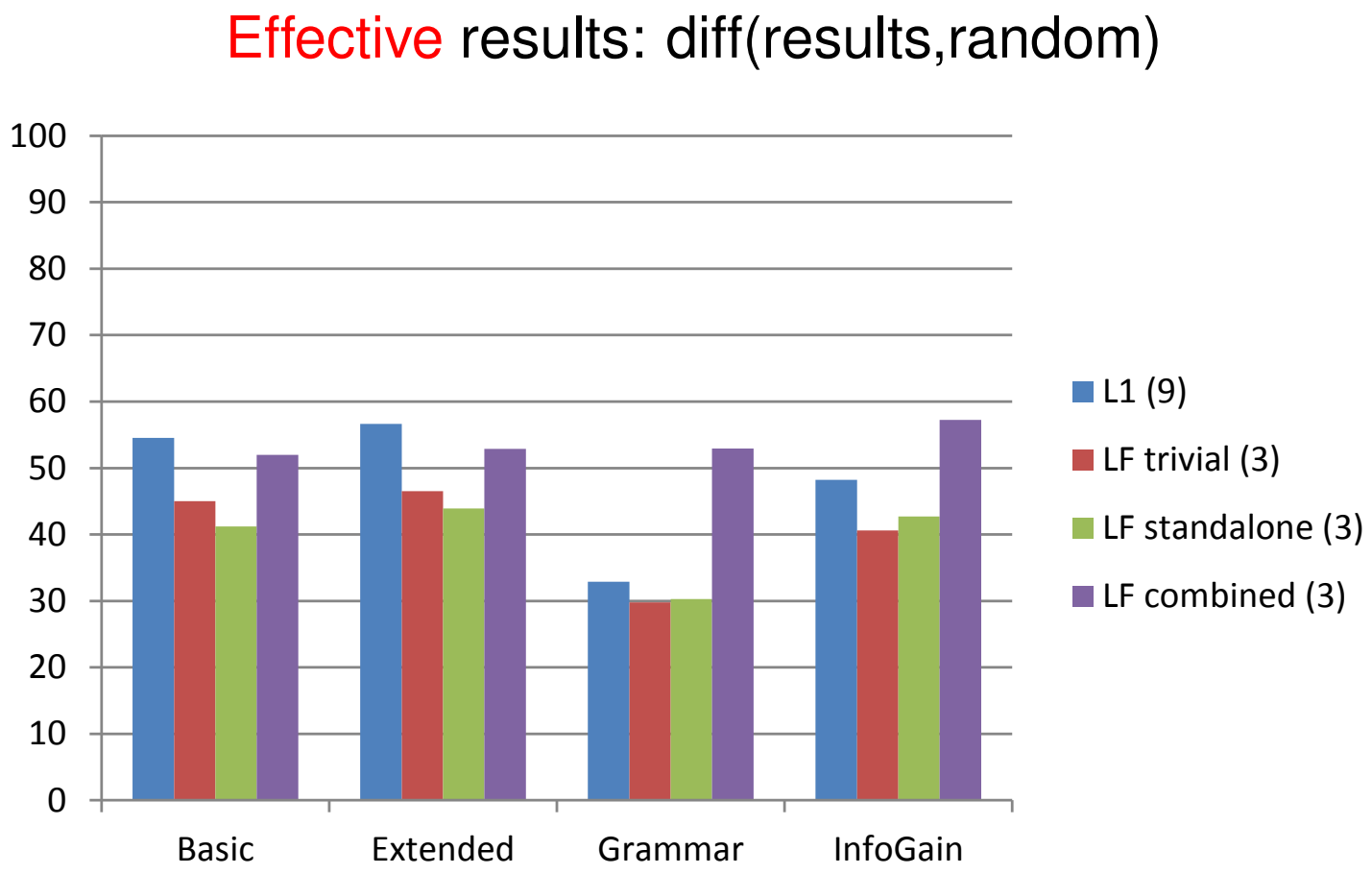


9-Class L1 vs. 3-Class LF – Results

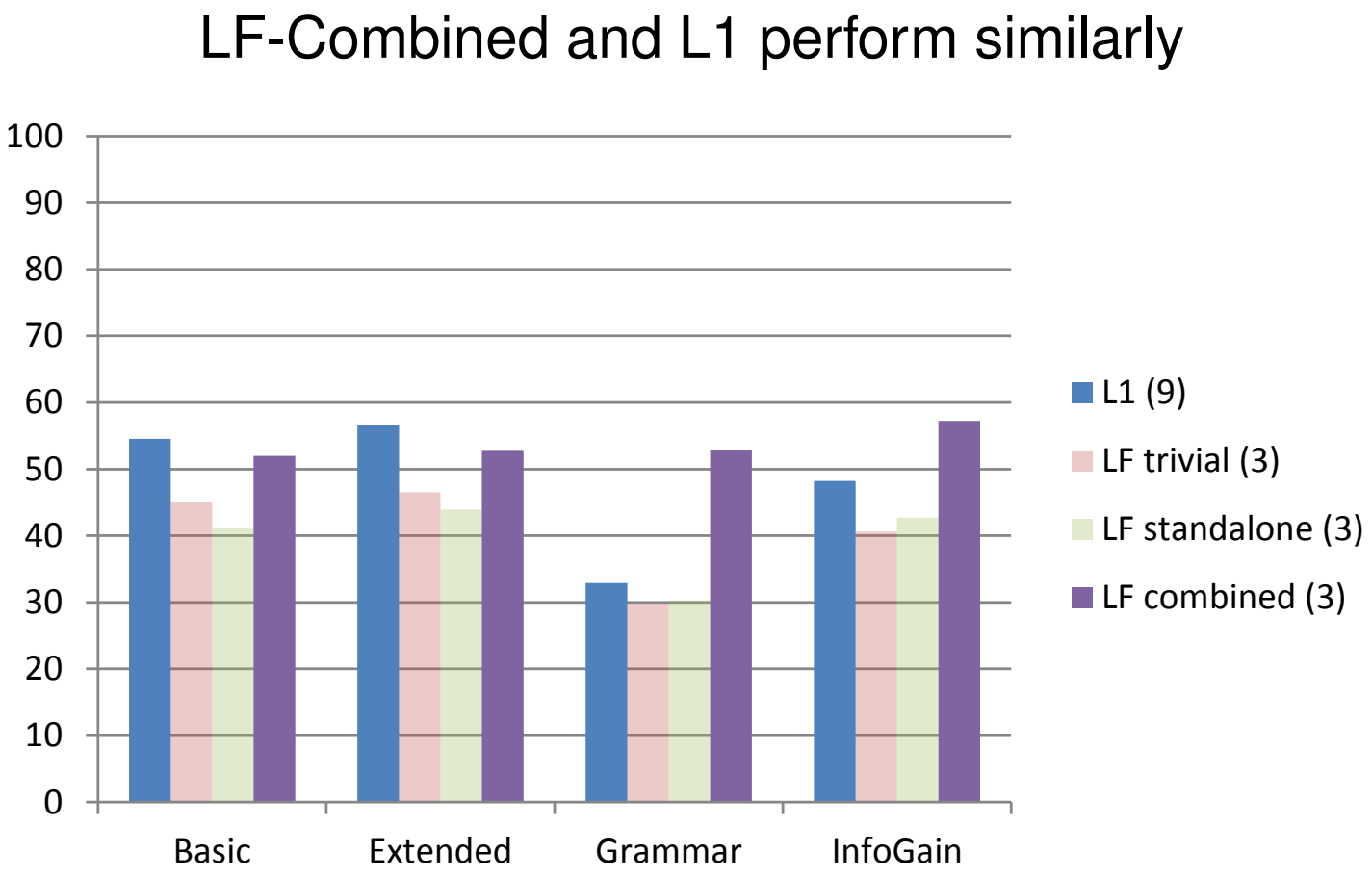
Combined outperforms Trivial, Standalone and L1



9-Class L1 vs. 3-Class LF – Effective Results



9-Class L1 vs. 3-Class LF – Effective Results



9-Class L1, Reclassify by LF

- Questions
  - Improve L1-ID with LF-ID?
- Correct L1-ID with LF information
  - Set confidence threshold t
  - Apply L1-ID, measure chosen L1 probability p
  - If $p \geq t$: take chosen L1
  - If $p < t$:
    - Apply LF-ID by Standalone / Trivial / Random
    - Reapply L1-ID only among languages in chosen LF
    - Take chosen L1

9-Class L1, Reclassify by LF

- Questions
  - Improve L1-ID with LF-ID?
- Correct L1-ID with LF information
  - Set confidence threshold t
  - Apply L1-ID, measure chosen L1 probability p
  - If $p \geq t$: take chosen L1
  - If $p < t$:
    - Apply LF-ID by Standalone / Trivial / Random
    - Reapply L1-ID only among languages in chosen LF
    - Take chosen L1

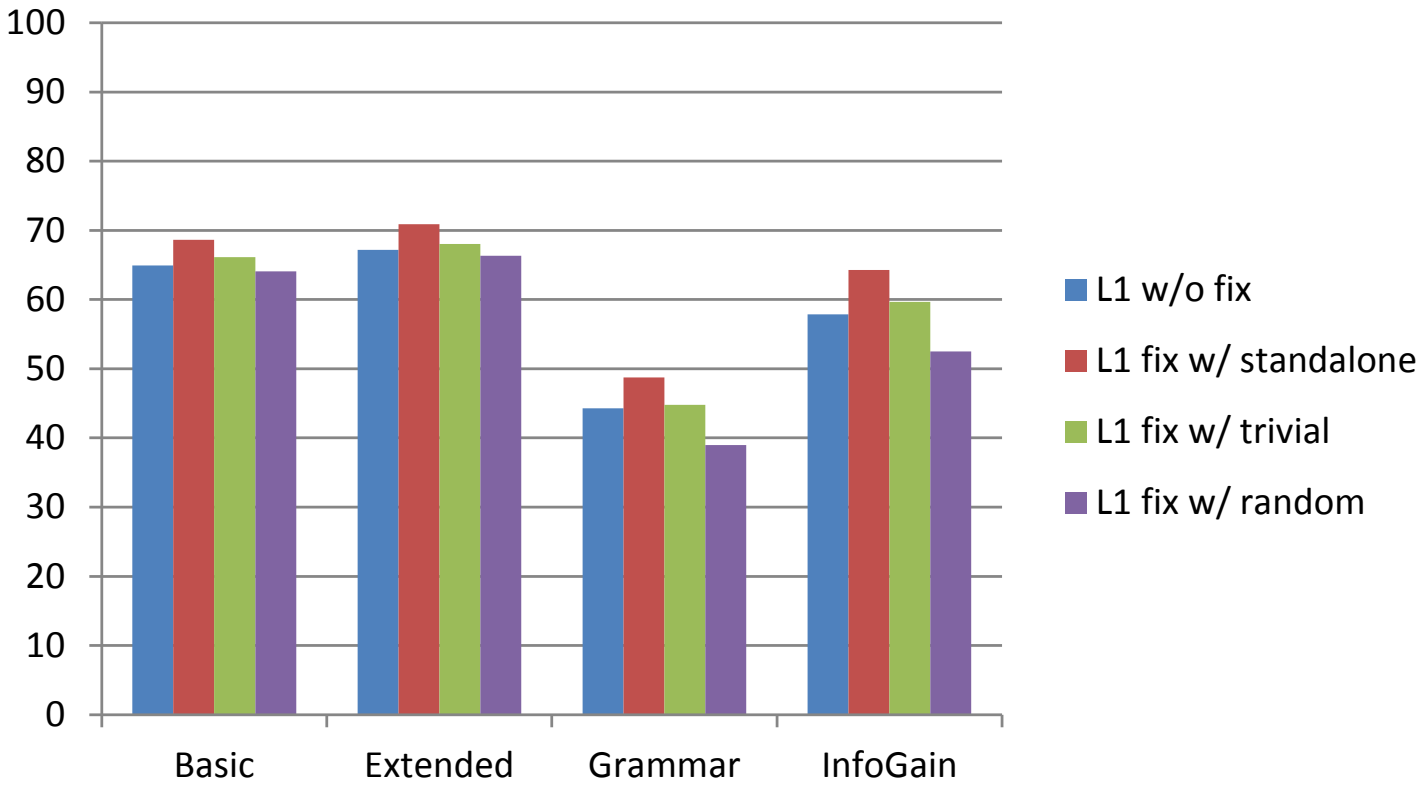
9-Class L1, Reclassify by LF

- Questions
  - Improve L1-ID with LF-ID?
- Correct L1-ID with LF information
  - Set confidence threshold t
  - Apply L1-ID, measure chosen L1 probability p
  - If $p \geq t$: take chosen L1
  - If $p < t$:
    - Apply LF-ID by Standalone / Trivial / Random
    - Reapply L1-ID only among languages in chosen LF
    - Take chosen L1

9-Class L1, Reclassify by LF

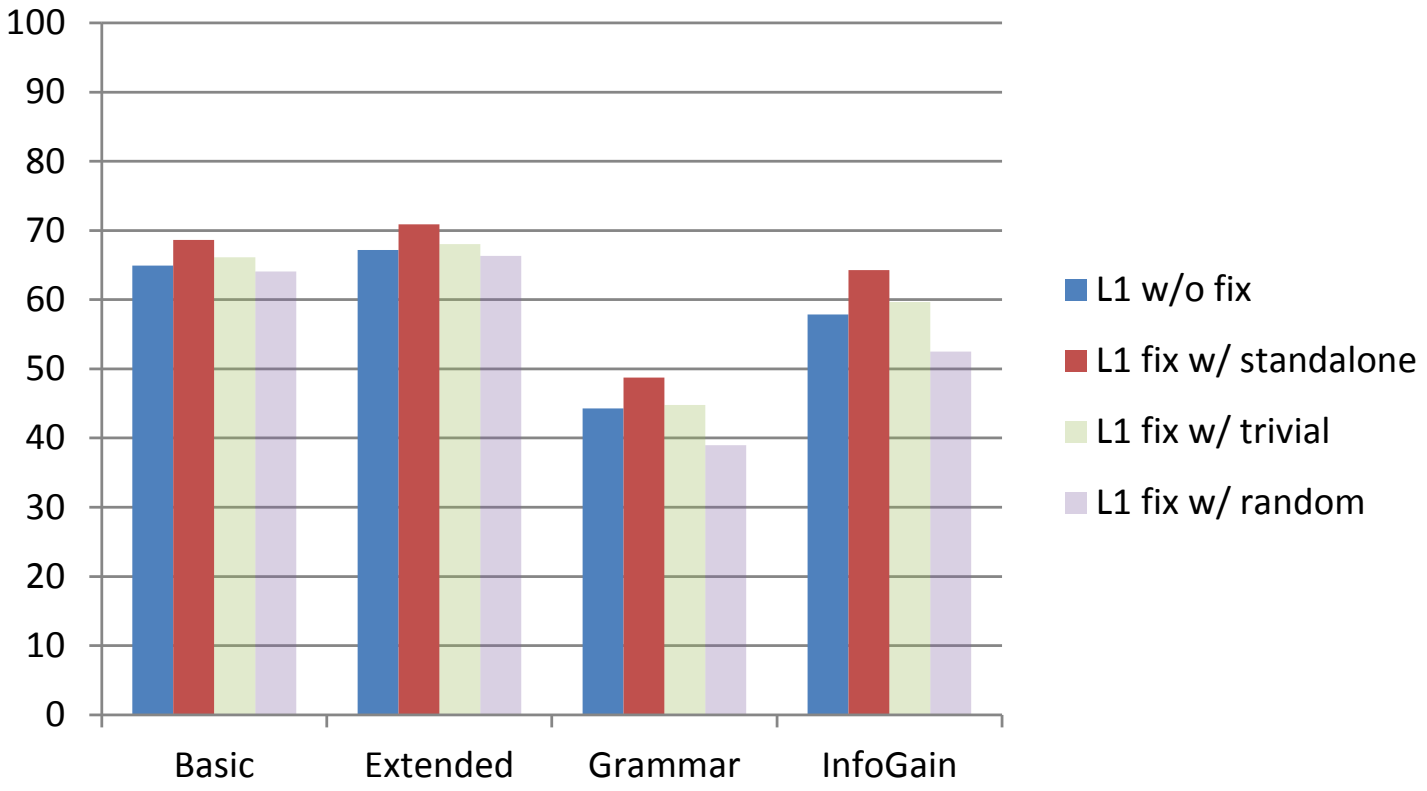
- Questions
  - Improve L1-ID with LF-ID?
- Correct L1-ID with LF information
  - Set confidence threshold t
  - Apply L1-ID, measure chosen L1 probability p
  - If $p \geq t$: take chosen L1
  - If $p < t$:
    - Apply LF-ID by Standalone / Trivial / Random
    - Reapply L1-ID only among languages in chosen LF
    - Take chosen L1

9-Class L1, Reclassify by LF – Results



9-Class L1, Reclassify by LF – Results

3.67%-6.43% increase in TPR using Standalone correction



Conclusion

- L1-ID can be improved by LF generalization step
  - 3.67%-6.43% increase in TPR for 9-Class L1 ($p < 0.01$)
- Adding grammar features is good for L1-ID and LF-ID
  - According to InfoGain, especially for LF-ID
  - Lexical features better for L1-ID than LF-ID
- Confident in L1 $\rightarrow$ Confident in LF of L1
  - Otherwise, find LF in a standalone experiment

Conclusion

- L1-ID can be improved by LF generalization step
  - 3.67%-6.43% increase in TPR for 9-Class L1 ($p < 0.01$)
- Adding grammar features is good for L1-ID and LF-ID
  - According to InfoGain, especially for LF-ID
  - Lexical features better for L1-ID than LF-ID
- Confident in L1 $\rightarrow$ Confident in LF of L1
  - Otherwise, find LF in a standalone experiment

Conclusion

- L1-ID can be improved by LF generalization step
  - 3.67%-6.43% increase in TPR for 9-Class L1 ($p < 0.01$)
- Adding grammar features is good for L1-ID and LF-ID
  - According to InfoGain, especially for LF-ID
  - Lexical features better for L1-ID than LF-ID
- Confident in L1 $\rightarrow$ Confident in LF of L1
  - Otherwise, find LF in a standalone experiment

Thank You

Thank You!

Questions?

- ams573@cs.drexel.edu
- http://psal.cs.drexel.edu/
  - JStylo: authorship attribution framework
  - Anonymouth: authorship anonymization framework

For Further Reading I

Julian Brooke and Graeme Hirst.
Measuring interlanguage: Native language identification with l1-influence metrics.
In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

Ghinwa F. Choueiter, Geoffrey Zweig, and Patrick Nguyen.
An empirical study of automatic accent classification.
In *ICASSP*, pages 4265–4268, 2008.

Dominique Estival, Tanja Gaustad, Son B. Pham, Will Radford, and Ben Hutchinson.
Author profiling for english emails.
In *10th Conference of the Pacific Association for Computational Linguistics (PACLING 2007)*, pages 262–272, 2007.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon.
Automatically determining an anonymous author's native language.
In *Proceedings of the 2005 IEEE international conference on Intelligence and Security Informatics*, ISI'05, pages 209–217, Berlin, Heidelberg, 2005. Springer-Verlag.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon.
Determining an author's native language by mining a text for errors.
In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 624–628, New York, NY, USA, 2005. ACM.

For Further Reading II

Hans van Halteren.
Source language markers in europarl translations.
In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 937–944, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

Sze-Meng Jojo Wong, Mark Dras, and Mark Johnson.
Topic modeling for native language identification.
In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 115–124, Canberra, Australia, December 2011.

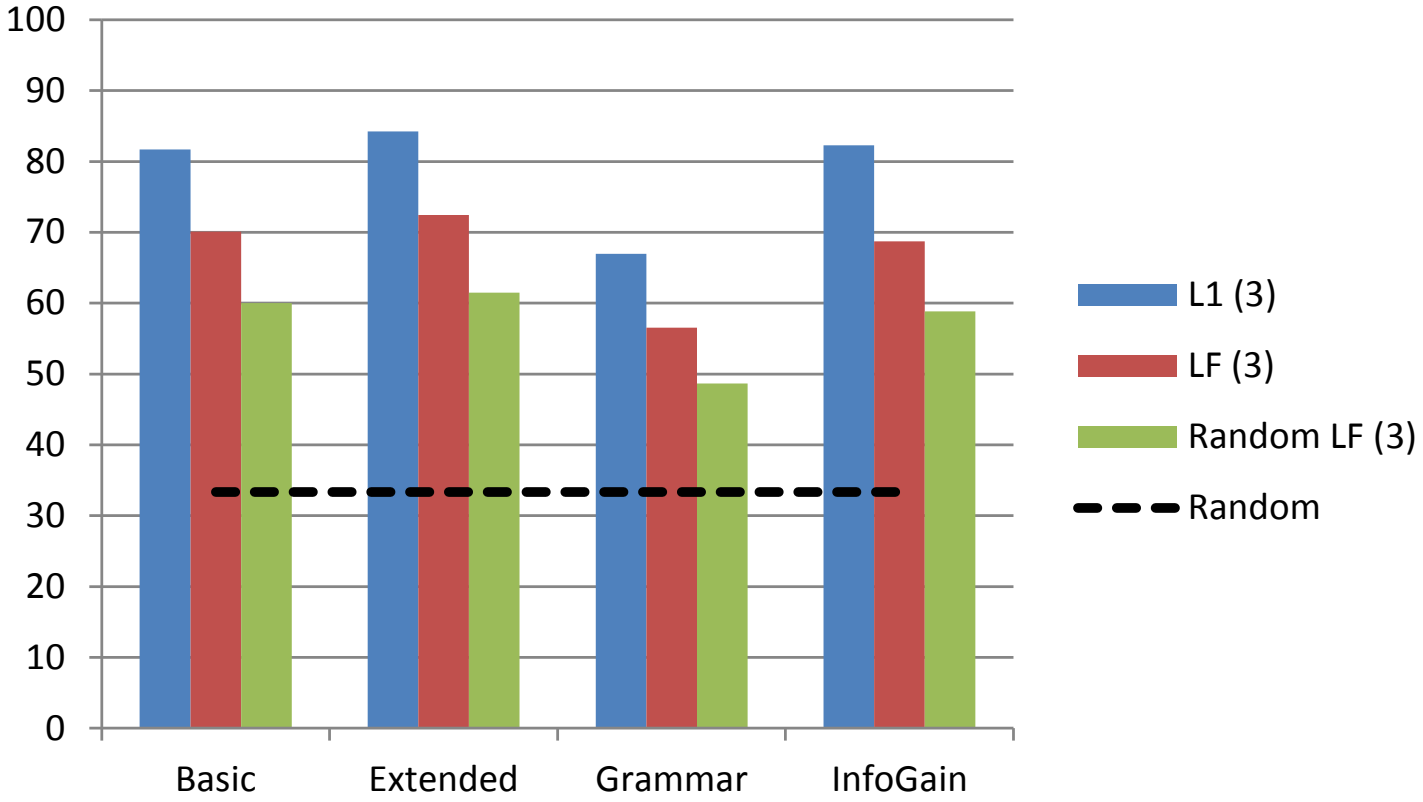
3-Class L1 vs. 3-Class LF

- Questions
  - L1-ID vs. LF-ID performance in same settings?
  - Real LF better than Random LF?
- Compared
  - 3-class L1-ID
  - 3-class LF-ID (*Standalone*), real LF
  - 3-class LF-ID (*Standalone*), randomly constructed LF

3-Class L1 vs. 3-Class LF

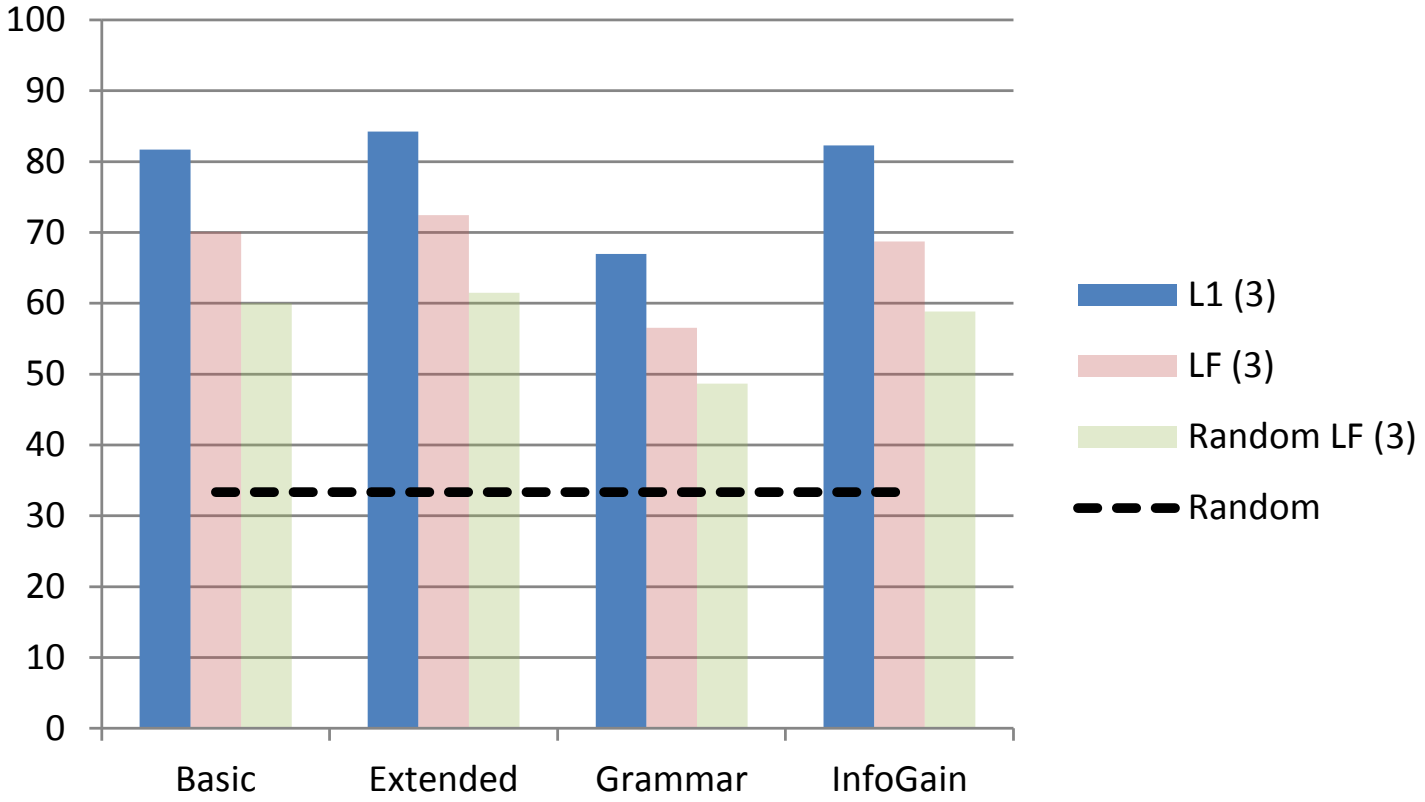
- Questions
  - L1-ID vs. LF-ID performance in same settings?
  - Real LF better than Random LF?
- Compared
  - 3-class L1-ID
  - 3-class LF-ID (*Standalone*), real LF
  - 3-class LF-ID (*Standalone*), randomly constructed LF

3-Class L1 vs. 3-Class LF – Results



3-Class L1 vs. 3-Class LF – Results

L1-ID more accurate than LF-ID in same settings



3-Class L1 vs. 3-Class LF – Results