D8 An Overview of Computational Linguistic Techniques for Forensic Purposes

Carole E. Chaski, PhD, ALIAS Technology, LLC, Georgetown, DE 19947*

Learning Overview: After attending this presentation, attendees will be familiar with techniques used in computational linguistics projects, such as Google®, and how these techniques can be useful for forensic purposes.

Impact on the Forensic Science Community: This presentation will impact the forensic science community by introducing the forensic uses of computational linguistics.

By its very nature, forensic science borrows from non-forensic science; the strongest forensic sciences, such as DNA identification and toxicology, were all respectable sciences before they were discovered to have profound use within investigations and thus became forensic DNA analysis and forensic toxicology (i.e., forensic sciences). This study presents computational linguistics as another non-forensic science that has forensic applications. Computational linguistics for forensic issues is relatively novel and is not represented in most crime laboratories, but it has been steadily growing since it was first introduced as a research paradigm in 1994, with results first reported in 1997.[1] Further, testimony using computational linguistics has been admitted under both *Daubert* and *Frye* standards.

Computational linguistics is the branch of linguistics that uses computer programs and statistical analysis to test theories of language structure and use. Linguistic theories create grammars of a language; computational linguistics uses this grammar to create software (called a parser) that analyzes the language, thus testing how well the grammar actual works on real data. Computational linguistics software has two functions: Type 1—a computer performs language-based tasks for a human; or Type 2—a computer helps humans use language and language-based analysis to perform other tasks. In tasks of type 1, the input is language and the output is still language. In tasks of type 2, the input is language and the output is something about the language, outside of language itself. The first tasks are meta-linguistic, (language about language) while the second tasks are extra-linguistic (language for something else).

Using Language-Based Software to Obtain Linguistic Information: The most famous example of type 1, metalinguistic tasks, is the search engine Google®. The search engine is run by a computer program that enables the machine to receive language as keywords, find and retrieve information, and output language in the form of a document set matching the keywords. In addition to search and information retrieval, other computational linguistics techniques include similarity detection (ranking most closely similar documents in a set), automatic summarization of documents, text classification, identification of the language(s) in a document, and machine translation.[2-7] For jurisprudence, both search and summarization can be useful in e-discovery, while similarity detection can be used for contract and motion reviews. For forensic digital examination, similarity detection can be used to find similar strings in assembly language, software languages in malware, and natural languages, while language identification can be useful in the examination of variable names in malware.[8] For criminalistics, text classification can be very useful in determining if a document is really the type of text that it seems to be (such as threats and suicide notes).[9]

Using language-based software to get extra-linguistic information: The type of computational linguistics in type 2 takes in language and outputs some fact or assessment related to the language behavior but outside language. Language analysis by machine can be used to automate personality assessment and to evaluate language for detection of mental disorders, such as delusional thinking, schizophrenia, and depression.[10,11] For forensic psychology, this type of software has an immediate use of presenting a quantitative “no skin in the game” evaluation with known error rates. Computational linguistics can automate credibility assessment or indications of deception in language; this possibility has immediate uses in forensic interviewing and interrogation.[12] Computational linguistics can be used to determine who authored a document.[13-15] For forensic document examination, determining the authorship can be useful in advancing a theory of case; automated tools also provide a known error rate. Computational linguistics can be used to determine the social demographics reflected in the language of a document, even if the author is unknown.[16,17] For criminalistics, narrowing the pool of potential suspects through social demographics can be a useful asset to an investigation. Further, automatic identification of terroristic messages helps criminal investigators monitor social media.[18,19]

Reference(s):

1. Chaski, C.E. Who Wrote It? Steps Toward a Science of Authorship Identification. *National Institute of Justice Journal.* September 1997. Also available through National Criminal Justice Reference Service: http://www.ncjrs.org NCJ 184604.
2. R. Baeza-Yates and B. Ribeiro-Neto. (2nd Ed). *Modern Information Retrieval: The Concepts and Technology behind Search.* Addison-Wesley, UK, 2011.
3. S. Brin, J. Davis, and H. Garcia-Molina. Copy Detection Mechanisms for Digital Documents. *ACM SIGMOD Record*, Vol. 24, ACM, 1995, pp. 398-409.
4. A. Nenkova and K. McKeown. A Survey of Text Summarization Techniques. In C.C. Aggarwal and C.X. Zhai (eds.), *Mining Text Data*, New York: Springer Science + Business Media, LLC 2012, pp. 43-76. doi: 10.1007/978-1-4614-3223-4_3.
5. C.C. Aggarwal and C.X. Zhai. A Survey of Text Classification Algorithms. In C.C. Aggarwal and C.X. Zhai (eds.), *Mining Text Data*, New York: Springer Science + Business Media, LLC 2012, pp. 163-222. doi: 10.1007/978-1-4614-3223-4_6.
6. R.D. Lins and P. Goncalves. Automatic Language Identification of Written Texts. 2004. In *Proceedings of the 2004 ACM Symposium on Applied Computing*. Pp. 1128-1133.
7. J. Hutchins. Retrospect and Prospect in Computer-Based Translation. 1999. Machine Translation Summit VII, 13th-17th September 1999, Kent Ridge Labs, Singapore. *Proceedings of MT Summit VII “MT in the great translation era.”* (Tokyo: AAMT), pp.30-44.

Copyright 2019 by the AAFS. Permission to reprint, publish, or otherwise reproduce such material in any form other than photocopying must be obtained by the AAFS.

[8] G. Frantzeskou, E. Stamatatos, S. Gritzalis, Chaski, C.E., and Howald, B. Identifying Authorship by Byte-Level N-Grams: The Source Code Author Profile (SCAP) Method. 2007. *International Journal of Digital Evidence*, Spring 2007, Volume 6:1.

[9] C.E. Chaski and D. Huddle. Is This a Real Suicide Note? Authentication Using Statistical Classifiers and Computational Linguistics. 2011. *Proceedings of the American Academy of Forensic Sciences*, 63rd Annual Scientific Meeting, Chicago, IL. 2011..

[10] J.W. Pennebaker, R.L. Boyd, K. Jordan, and K. Blackburn. (2015). *The Development and Psychometric Properties of LIWC2015*. Austin, TX: University of Texas at Austin.

[11] S.A. Khan. (2015). *Screening and Investigating Creative Writing of Emotionally Disturbed Pakistani University Students: Linguistic Preventive Intervention*. PhD Dissertation. University of Azad Jammu Kashmir, Muzaffarabad, Pakistan.

[12] C.E. Chaski, A. Almela, G. Holness, and L. Barksdale. (2015). WISER: Automatically Classifying Written Statements As True or False. *Proceedings of the American Academy of Forensic Sciences*, 67th Annual Scientific Meeting, Orlando, FL. 2015.

[13] C.E. Chaski. Who's At the Keyboard? Recent Results in Authorship Attribution. 2005. *International Journal of Digital Evidence*. Volume 4:1. Spring 2005. Available at http://www.ijde.org.

[14] C.E Chaski. The Keyboard Dilemma and Author Identification. 2007. In Sujeet Shinoi and Philip Craiger, editors. *Advances in Digital Forensics III*. New York: Springer.

[15] A. Rocha, W.J. Scheirer, C.W. Forstall, T. Cavalcanter, A. Theophilo, B. Shen, A.R.B. Carvalho, and E. Stamatatos. (2016) Authorship Attribution for Social Media Forensics. *IEEE Transactions on Information Forensics and Security*.

[16] Schler, J., Koppel, M., Argamon, S. and Pennebaker, J. Effects of Age and Gender on Blogging. *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*. (2006) pp. 199-205.

[17] H.A. Schwartz, J.C. Eichstaedt, M.L. Kern, L. Dziurzynski, S.M. Ramones, M. Agrawal M., et al. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE* 8(9): e73791. https://doi.org/10.1371/journal.pone.0073791.

[18] M. Ashcroft, A. Fisher, L. Kaati, E. Omer, and N. Prucha. (2015). Detecting Jihadist Messages on Twitter. *Intelligence and Security Informatics Conference*, (EISIC). 10.1109/EISIC.2015.27.

[19] A.H. Johnston and G.M. Weiss (2017). Identifying Sunni Extremist Propaganda With Deep Learning. *Computational Intelligence* (SSCI) 2017 IEEE Symposium Series on, pp. 1-6, 2017.

Computational Linguistics, Forensic Science, Extra-Linguistic Investigation

Copyright 2019 by the AAFS. Permission to reprint, publish, or otherwise reproduce such material in any form other than photocopying must be obtained by the AAFS.