From Shakespeare to Twitter: What are Language Styles all about?

Wei Xu
Department of Computer Science and Engineering
The Ohio State University
weixu@cse.ohio-state.edu

Abstract

As natural language processing research is growing and largely driven by the availability of data, we expanded research from news and small-scale dialog corpora to web and social media. User-generated data and crowdsourcing opened the door for investigating human language of various styles with more statistical power and real-world applications. In this position/survey paper, I will review and discuss seven language styles that I believe to be important and interesting to study: influential work in the past, challenges at the present, and potential impact for the future.

1 Top Three Problems

The top three problems for studying language styles are data, data and data. More specifically, they are data shortage, data fusion, and data annotation problems. The data shortage problem has been improving, which is the main reason that there is surge in the number of research studies on language styles. The data fusion problem is more specific to the area, due to the subtle and often subjective nature of linguistic styles. For instance, while men and women talk in different ways (note this is not the same as talking about different things), they also talk about a lot of things in an indistinguishable way. Moreover, there is also a huge variance from one man to another, one woman to another. The styles are often fused together in the data and not easy to separate out or make black-and-white judgements on. This also leads to challenges in data annotation or data collection, comparing to other NLP tasks (e.g. question answering). Throughout the rest of this paper, we shall see many creative solutions, interesting work, and promising potential.

2 Seven Styles of Language

Disclaimers: (i) We discuss primarily in the context of natural language processing research; (ii) There are certainly more than seven language styles as there are more than seven wonders in the world.

2.1 Simple and Short

Text simplification is one of the earliest topics in computational linguistics that directly deals with language styles, rewriting regular texts into simpler versions for people with limited reading capabilities. The major transition from rule-based to machine learning approach for automatic sentence simplification did not happen until 2010 after Simple English Wikipedia became available. It is worth noting that the Simple Wikipedia data has some issues on the quality and degree of simplicity (Xu et al., 2015b). The shortage of high quality data is becoming gradually alleviated as the Newsela corpus (Xu et al., 2015b) of professionally edited 1000+ articles is released, and as more and more attention and appreciation are given by the research community to data construction (Brunato et al., 2016; Hwang et al., 2015). Multiple studies have shown crowdsourcing workers can produce high quality simplifications (Xu et al., 2016; Amancio and Specia, 2014; Pellow and Eskenazi, 2014), though it is costly to scale up. Data will remain a central problem as the data-hungry neural generation models (Nisioi et al., 2017) are a promising direction for future work.

Besides data, another severe problem is evaluation. In fact, one common human evaluation that uses a five point Likert scale on grammaticality, meaning and simplicity should be considered
unacceptable when deletion is involved, as it unfairly biases towards deletions over paraphrasing. There has been some progress on creating automatic evaluation metrics (Xu et al., 2016) and exploring new human evaluation methodologies (Xu et al., 2016; Nisioi et al., 2017; Siddharthan and Mandya, 2014). We are going to need more data, clever ideas and careful evaluation designs.

For the record, everything about sentence simplification is much harder than sentence compression\(^2\) primarily due to the interactions between deletion and paraphrasing. Like simplification, previously, sentence compression also use human evaluation with Likert scale on grammaticality and meaning. However, it is shown to be problematic without controlling for compression ratio (Napoles et al., 2011). Now sentence compression systems are mostly compared at the same compression ratio. It is also worth noting that neural compression is similarly lacking in large-scale parallel data (Toutanova et al., 2016) and currently relies on news headline data which results in headline-like outputs (Filippova et al., 2015; Rush et al., 2015).

\(^2\)which is closely related to, sometimes used interchangeably with, though different from, abstractive summarization, headline generation, sentence fusion.

2.2 Instructional and Robotic

Despite the fact that instructional language is important in our everyday lives, there have been relatively limited efforts to design automated algorithms that link language to action in real world applications. Largely because of the limited availability of annotated datasets which are much-needed for training and evaluating machine learning models, existing works are primarily on cooking recipes (Tasse and Smith, 2008), airline booking conversations (Zettlemoyer and Collins, 2007), software help documents (Branavan et al., 2009) and robot navigation commands (Chen and Mooney, 2011). In particular, cooking recipe has sprouted a rich line of research as a proxy to robotic instructions (Bollini et al., 2013; Jermisurawong and Habash, 2015; Kiddon et al., 2015). Recent efforts aim to study natural language instructions for biology lab experiments (Kulkarni et al., 2017). Two closely relevant research areas, semantic parsing and dialog, have also both made major advances in recent years to utilize large-scale data via weak supervision (Cai and Yates, 2013; Artzi and Zettlemoyer, 2013) and neural network models (Lee et al., 2016; Misra and Artzi, 2016). The 1st Workshop on Language Grounding for Robotics (RoboNLP) will be held at ACL 2017. We shall expect research on instructional language become more and more fruitful in the near future.

2.3 Historical and Evolving

The rise of digital humanities certainly helps to provide more digitized materials for learning techniques. Historical documents are proven fun (in the other word, hard) to work with. Garrette and Alpert-Abrams (2016) used the following example to present the challenges of having multiple unknown fonts and inking on a single page of a book in the Primeros Libros corpus:
(Plank, 2018). The most apparent case is out-of-vocabulary (OOV) words (van der Wees et al., 2015; Seraj et al., 2015), especially new emerging named entities and newly coined words (e.g. “selfie”, “Brexit”). This problem will become more pressing and more feasible to study as more and more time-sensitive online text data is accumulating. Learning up-to-date paraphrases (Lan et al., 2017), vector semantics (Cherry and Guo, 2015) and character-based neural models (Ling et al., 2015; Rei et al., 2016) from online data streams could be plausible solutions that connect unseen data with known expressions.

2.4 Colloquial and Internet

As social media started booming, especially after Twitter released the streaming API for free in 2010 that provides real-time tweets as posted, there is a huge explosion on social media research. Multiple workshops are dedicated to this special type of text including the Workshop on Noisy User-generated Text (WNUT) and Workshop on Making Sense of Microposts (#microposts) that hold annual shared tasks. Before that, most unedited text data (vs. well-edited such as news) is from web forums and blogs, while short message service (SMS) and email data are limited to rather small amounts due to privacy reasons (Baldwin et al., 2013). Interesting research falls into two camps: normalize lexical variants to standard form (Han and Baldwin, 2011; Xu et al., 2013) or develop domain adapted NLP systems (Ritter et al., 2011; Gimpel et al., 2011; Kong et al., 2014; Tabassum et al., 2016). The iconic opinion paper What to do about bad language on the Internet by Jacob Eisenstein (2013) highlighted this divide.

There is a third point we have often missed. Besides the noisy hard-to-understand Internet language, many users also use rather standard language on social networks, formal or colloquial. Don’t forget that all the traditional news agencies also have Twitter accounts (Hu et al., 2013). Can we make the connections between the formal and colloquial languages as they are heavily mixed on social media? I think the answer is yes, and the twin research topics of paraphrasing and semantic similarity could be part of the solution as many language styles are heavily mixed on social media. For example, in the SemEval shared task PIT-2015 corpus (Xu et al., 2015a), the figurative meaning of the phrase “on fire” is captured by the sentential paraphrase of “Aaaaaaaaand stephen curry is on fire” and “What a incredible performance from Stephen Curry”. Semantic equivalences, as formal as “fetuses” and “fetal tissue” (Lan et al., 2017) or as informal as “gets the boot from” and “has been sacked by” (Xu et al., 2014; Xu, 2014), can also be learned automatically from Twitter data. Not to mention that there are also studies that focus on multiword expressions (Schneider and Smith, 2015), idioms (Muzny and Zettlemoyer, 2013), and slang.

2.5 Gendered and Personalized

One unique and exciting opportunity offered by social media data is to learn about the users authoring the texts. Much interesting research on gender difference in language styles appeared in the past few years. Besides gender (Verhoeven et al., 2016; Bamman et al., 2014), other user attributes such as age (Sap et al., 2014), race (Jørgensen et al., 2015) and personality (Schwartz et al., 2013; Ruan et al., 2016; Plank and Hovy, 2015) are also commonly studied for social science and strongly motivated by commercial usages of profiling users and personalized services. Leveraging user demographic factors also shows benefits on improving natural language processing applications such as sentiment analysis (Volkova et al., 2013) and sarcasm detection (Bamman and Smith, 2015).

One particularly interesting challenge is how to handle the situation that stylistic differences (e.g. female users more likely use “wonderful” while male users use “superb”) are much more subtle than topical preferences (e.g. using word “husband” is a strong indicator of female user). Our recent work (Preotiuc-Pietro et al., 2016) isolated stylistic differences from topic bias by using paraphrase pairs and clusters, and showed their predictive power in user profiling and potential for future work. We also found crowdsourcing workers are surprisingly good at perceiving gender from lexical choices when aggregating their judgments – an infamous phenomenon of so-called The Wisdom of Crowds (Surowiecki, 2005). Beyond lexical choice, Johannsen et al. (2015) further showed demographic differences in syntactic variances using multilingual data of online customer reviews and universal dependency parsing.

Although unrelated to linguistic styles, the readers may find He Said, She Said: Gender in the ACL Anthology (Vogel and Jurafsky, 2012), a paper on gender-based statistics of NLP researchers, interesting.
Another subsequent challenge is how to transfer the subtle style differences into natural language generation and dialog systems. While we were able to transform contemporary texts into Shakespeare style (Xu et al., 2012b), we found gendered language style much harder to impose. It is possibly that because we have not found the right data for evaluation, for instance, it is hard to expect a randomly drawn sentence to be possible to take on a feminine or masculine style. It could also be the case that it is easier for finer-grained language style to show distinctions. One evident example is author recognition based on an individual’s frequent word choices (Clark and Hannon, 2007). Another example is persona-based dialog system that not only captures background knowledge of a user (Li et al., 2016) but also speaking style (Mizukami et al., 2015). It is not a coincident that the later work (Mizukami et al., 2015) is on spoken Japanese, which exhibits extensive gender differences as well as honorifics (not as much in written Japanese).

2.6 Pervasive and Framing

The increasing availability of data also make feasible to study the textural characteristics of persuasion, argumentation and framing in realistic (not laboratory) settings and quantitatively. Besides movie quotes, political speeches, and tweets (Guerini et al., 2015), many interesting data are created and discovered, leading to a growing number of studies. Online discussion platforms provide almost ideal real world data with users stating, reasoning and contesting opinions (Somasundaran and Wiebe, 2009), and sometimes even with explicitly marked successful arguments such as ChangeMyView on Reddit. One recent work (Tan et al., 2016) found that in the ChangeMyView data, after controlled for similar arguments, stylistic choices in how the opinion is expressed carry more predictive power on how likely a user to be persuaded than how likely an argument is persuasive. However, predicting persuasiveness turns out to a difficulty task with about 60-65% accuracy using bag-of-words and linguistic features, in contrast of 75-85% accuracy for predicting politeness. Another interesting work (Recasens et al., 2013) utilized Wikipedia edit history to study biased language (e.g. “stated” vs. “claimed”) as well as framing (e.g. “pro-life” vs. “anti-abortion”). The recent construction of the Media Frames Corpus (Card et al., 2015) presents another encouraging opportunity to study framing. The legal domain, such as supreme court documents, is another common place for arguments (Sim et al., 2015) and would possibly be used for studying linguistic styles.

2.7 Polite and Abusive

Another angle that has been looked at is the politeness conveyed in language. Unlike many other styles that come in close pairs (e.g. formal vs. informal, feminine vs. masculine), the polite language does not necessarily have an impolite counterpart. In addition, politeness is expressed more through function words. For example, showing gratitude by “I appreciate that” or apologizing by “Sorry to bother you”. In fact, the phrase “in fact” can be negative as “in fact you did...”. Many other cues are identified and annotated (Danescu-Niculescu-Mizil et al., 2013) on the online interchanges of Wikipedia editors and StackExchange QA users, which can train classifier to predict politeness at about 80% accuracy. A recent study (Voigt et al., 2017) also used automatic methods to examine the respectfulness of police officers toward white and black people from transcripts of body-worn camera footage.

In other words, abusive language is closely related to politeness but not the reverse. The targets could vary from one swear word to multi-sentences, such as the mean tweet Barack Obama read on Jimmy Kimmel’s show: “Obama’s hair is looking grayer these days. Can’t imagine why since he doesn’t seem to be one bit worried about all that’s going on.” The context-dependent nature makes it challenging to collect data or design experiments. Moreover, although bullying traces are abundant, it is a tiny fraction out of random samples which is estimated to 0.02~0.73% of a 95% confidence internal on 2011 TREC Microblog track corpus (Xu et al., 2012a). The compromise is to look at tweets that include keywords “bully”, “bullied”, “bullying” instead, which is inspiring and an important first step, but far from satisfying. Another representative solution is a carefully designed crowdsourcing experiment which reveals patterns of Internet trolling behavior using user comments on CNN.com news website (Cheng et al., 2017). Perhaps, the 1st
Workshop on Abusive Language Online (ALW) at ACL 2017 will spark more ideas. I would like to quote an anonymous source who raised a thoughtful question: “Under what circumstances is language use considered to be an abuse? For example, in many states when a woman criticizes her husband in public, this might be considered there as abuse of language or hate speech”, as a reminder of being aware and mindful of the great social factors and impacts embedded in the research of language styles.

3 Conclusion

At this point of the development, natural language processing research ranges a wide variety of genre, domain, register or type of data. I think the term style is an all-in-one umbrella concept to bring researchers and scattered attentions in various NLP subareas to a common place. There are certainly many nuances in language styles besides those mentioned in this paper. For example, connotation (e.g. “childlike” vs. “childish” vs. “youthful”) (Rashkin et al., 2016; Carpuat, 2015) and geographical lexical variations from regional (e.g. “sode” vs. “coke” vs. “pop”) to cross-country (e.g. Australian vs. American English) (Eisenstein et al., 2010; Garimella et al., 2016; Han et al., 2016). There are also certainly many other relevant works besides those mentioned in this paper. Last but not least, we would like to point out Dan Jurafsky’s recent book The Language of Food (2014) and one more paper: Do Linguistic Style and Readability of Scientific Abstracts Affect their Virality? (Guerini et al., 2012).

Acknowledgments

I would like to thank three anonymous reviewers for helpful feedback, Julian Brooke, Thamar Solario and Moshe Koppel for organizing the EMNLP 2017 Workshop on Stylistic Variation.

References


David Bamman and Noah A. Smith. 2015. Contextualized sarcasm detection on Twitter. In ICWSM.


Marine Carpuat. 2015. Connotation in translation. In Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis.


Barbara Plank and Dirk Hovy. 2015. Personality traits on Twitter -or- how to get 1,500 personality tests in a week. In Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis.


