Text Normalization in Noisy Channels

Hila Weisman, Peter Izsak, Inna Achlow, and Victor Shafran

NICE System
Zarhin 13, Ra'anana

Abstract

A major challenge in customer-oriented enterprises is conducting Multi-Channel Interaction Analytics: gaining insight from collaborated data sources. Today, end users interact with companies using websites, mobile applications, text messages and social media platforms. Moreover, agents themselves are required to document the call flow and content in a fast-paced call center surrounding. Within these channels there are varying degrees of creativity and domain-specific terms, incurring many typos, abbreviated forms of high-frequency, domain-specific words and product names. These lexical variations make it hard for statistical algorithms to collect sufficient and correct data, for categorization engines to identify relevant calls, for LVSCR to correctly identify OOV vs. erroneous words etc.

In this work we present a robust, fast and accurate text normalization solution which is able to deal with non-standard word forms and normalize them into their canonical form. We focus on two distinct noisy channels: Agent Notes and Feedback SMS. In Agent Notes, agents in call centers summarize the content of current calls from customers in order to have it later categorized and reviewed by their supervisors and business analysts. In Feedback SMS, a customer is answering a general question sent via SMS regarding the service she received. These channels are distinctly different from typical noisy channels, such as SMS messages between peers and Twitter messages, as our study shows. The uniqueness of these channels, in addition to the run-time constraints involving Real-Time Analytics, make it harder to use current state-of-the-art approaches [2, 3, 5, 1] which deal with different error types and are often computationally expensive.

We approach the task by first investigating and comparing the prevalent OOV types (abbreviations, spelling errors, typos, proper names etc) of each channel. Based on this investigation we build a new normalization model comprised of three main components: OOV detector, spell corrector and abbreviations corrector. We factor these components, while keeping the distinction between each channel, into a unified model which outputs the most probable correction for a given noisy word. Each such component is built

using minimal supervision, incorporating linguistic knowledge with Machine Learning methods such as Word2Vec [4] and Probabilistic Classifiers such as Naive Bayes [6]. Furthermore, we devise an efficient method to detect and correct noisy text as a two-step process where the heavy computation is performed as an offline preprocessing step, making the online step fast enough to be utilized in online scenarios of real-life Interaction Analytics applications.

We conduct evaluations of our model's performance in the different channels against state-of-the-art methods and perform thorough error analysis on the different OOV types and their correlation with actual un-canonical form frequencies. In addition, we evaluate our system's run-time performance against different data-sets of varying sizes and complexity.

References

[1] E. Brill and R. C. Moore. An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 286–293. Association for Computational Linguistics, 2000.

[2] P. Cook and S. Stevenson. An unsupervised model for text message normalization. In *Proceedings of the workshop on computational approaches to linguistic creativity*, pages 71–78, 2009.

[3] B. Han and T. Baldwin. Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 368–378, 2011.

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[5] Y. Park and R. J. Byrd. Hybrid text mining for finding abbreviations and their definitions. In *Proceedings of the 2001 conference on empirical methods in natural language processing*, pages 126–133.

[6] I. Rish. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, pages 41–46. IBM New York, 2001.