Social (distributed) language modeling, clustering and dialectometry

David Ellis
Facebook
Palo Alto, CA
dellis@facebook.com

Abstract

We present ongoing work in a scalable, distributed implementation of over 200 million individual language models, each capturing a single user's dialect in a given language (multilingual users have several models). These have a variety of practical applications, ranging from spam detection to speech recognition, and dialectometrical methods on the social graph. Users should be able to view any content in their language (even if it is spoken by a small population), and to browse our site with appropriately translated interface (automatically generated, for locales with little crowd-sourced community effort).

1 Introduction

We approach several key questions from a data-driven (statistical) perspective, drawing on large, dynamic annotated corpora:

1. What social factors affect language change (and evolution)? How?
2. How do individuals adjust their speech or writing depending on context and audience?
   (e.g., register, formality, humor, reference)
3. What are the minimum requirements for a language (or dialect)?
   (e.g., number of speakers, corpus size)
4. Is a common language necessary for communication?
   Can a pidgin be predicted from its speaker-population?

To this end, we describe a framework for language modeling on the social graph, which incorporates similarity clustering and lays the groundwork for personalized (and multimodal) machine translation.

2 Related Work

Research on large scale language modeling (Brants et al., 2007) has addressed sharding, smoothing and integration with a machine translation pipeline. Our work takes a similar approach, using Hadoop (Borthakur, 2007) and Hive to query and process distributed data. Social annotations enhanced smoothing for language modeling in the context of information retrieval (Xu et al., 2007), and hierarchical Bayesian networks were used (Zhou et al., 2008) to incorporate user domain interest in such models. Language models are often used to detect spam, including in social bookmarking (Bogers and van den Bosch, 2008).

Proposed scoring models for social search (Schenkel et al., 2008) use friendship strengths and an extension of term frequency[1]. These could benefit from a deeper integration with friends' language models, perhaps to approximate a user-specific inverse document frequency, rather than treat each tag by a user as equally relevant to all his friends of a given (relationship) strength. Strehl et al. (2000) found that similarity clustering perform best using weighted graph partitioning.

3 Language Model

An individual's language model is a mixture of their locale (or another language they speak) and token frequencies from the content they produce (write) and consume (read). Since we have hundreds of milliions of users, each of whose language model can depend on a variety of data sources, it is essential to distribute these counts (and other figures derived from them) in a way that optimizes the efficiency of our access patterns[2].

We also tried clustering users, and representing the language of each as deviations from its neighbors (or the norm of the cluster). However,

[1] Called "socially-enhanced tag frequency".
[2] See Section 5 for discussion of a variety of use cases.

Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing, ACL-IJCNLP 2009, pages 1–4,
Suntec, Singapore, 7 August 2009.

there are significantly more edges than nodes in our graph (more friendships than people), so this alternative is less efficient.

An individual's language use varies greatly depending on his interlocutor or audience[3]. Messages I send (privately) to a friend differ in style from comments I make on a public photo of my nephew, which in turn differ from my writing style as realized in an academic or industry paper or article.

An obvious optimization is to describe a minimum spanning tree (MST) on the graph, where each edge is weighted according to the similarity of dialects associated with the nodes (individuals, groups or other entities) it connects. Then, language models of nodes connected by the MST can depend on each other's counts. Singletons default to the general language model from their locale.

3.1 Detecting Deviations

People who aren't friends (and have no mutual friends or other evident connection) may yet use more similar language than siblings. This example seems highly improbable or unnatural, and in fact serves as a good heuristic for detecting compromised, spam-sending accounts (even if not organized in a botnet).

If a user sends a message with high perplexity:

1. Their account is compromised, and being used to spam (or phish) their friends.
2. They are using a different language than usual. Users are often bilingual (sometimes multi)-, so we may not yet have realized they are proficient in a given language.
3. There may be a problem with the language model:
   (a) large vocabulary (tends to inflate perplexity)
   (b) genre mix (user interface v. user communication)

3.2 Locale Induction

A regional cluster of personal language models can be combined to create a new locale. A crowdsourced translation process (Ellis, 2009) can thus be bootstrapped by indirect community contributions.

4 Machine Translation

For an English-sepaking user, in order to optimize the probability of the target (translated) sentence given its source (Foreign), we follow Och and Ney's (2004) optimization of a set of feature functions:

$$\hat{\boldsymbol{e}} = \arg\max_{\boldsymbol{e}} \sum_{m=1}^{M} \lambda_m h_m(\boldsymbol{e}, \boldsymbol{f})$$

It is thus easy for us to aggregate scores from multiple language models (e.g., from individuals comprising your network of friends or others you interact with).

Our distributed, individual language models can be a component of personalized machine translation, where the target language may be a penpal's. Either the decoder incorporates the counts from user communications by supplementing the language model used in its n-best candidate search, or it uses the locale's general language model and factors in individual variance in a rescoring step.

We plan to offer inline statistical machine translation (SMT) of user-generated content, where the translation model combines features from:

1. Our (interface) translations corpus for the language pair
2. Related langauges or dialects[4]
3. Linguistic rules (Ellis, 2009), in some combination of:
   (a) Explicitly encoded
   (b) Induced from training corpora
   (c) Borrowed from related languages (esp. for relatively minor or resource-poor)

4.1 Sparse Data

Data sparseness is clearly an issue for modeling with this degree of specificity, so we explore a range of possible smoothing techniques, as well as methods for leveraging resources from related languages (Genzel, 2005). If a user signed up for Facebook last week, (s)he may not yet have connected with many friends or shared much content (which exacerbates the problem).

[3] This is not novel in or of itself, but the scale of our data and experiments should lead to finer-grained understanding, both of issues peculiar to a single language or its family, and of language universals (or.patterns; priors likely intuitively encoded).

[4] e.g. Spanish (Argentina, Spain), Chinese (Mandarin, Cantonese (Hong Kong, Taiwan)), or Finnish and its neighbors: inc. Estonian, Sámi, Komi

Domain adaptation is also important, since the base corpus is for a user interface: usually more formal, less varied than conversation. Ideally, we would like to capture not only language change (diversion, creolization) but an individual's linguistic evolution in a variety of contexts:

- She learns a language, practices its use, becomes increasingly accustomed to its twists and turns (syntactic, lexical, morphological, etc.)
- His mood shifts, he moves into a new apartment or city, let alone grander (potentially dynamic) features of context
- A startup company is suddenly more visible (e.g., resulting from press coverage, or a tech blogger's reference), and so an image (and website design, copy) revamp is in order.
- Afflicted with post-traumatic stress, after sensory deprivation, or in cases of neurological disorders or brain damage.

5 Similarity

We use a pipeline to cluster strings (to suggest translations) and users (based on language use):

1. Preprocessing
   - normalization (lowercasing)
   - {segment,{lemmat,token}iz}ation
2. Similarity (pick one)
   - fuzzy (hash) similarity[5]
   - string edit distance
   - phonetic (or phonological) edit distance
   - language model perplexity
   - KL-divergence (btn. language models)
3. Clustering (modular: select-an-algo)
   - hierarchical (agglomerative or divisive)
   - K-means (partitioning)
   - graph-theoretic methods (cover as opposed to cluster)

This is architected for ease of experimentation and modification, testing and tuning, so any combination of the above should be functional. Some applications of similarity require high accuracy but can be processed offline, whereas others need to be computed in less than ten milliseconds in response to a live query.

[5]i.e., Jaccard coefficient (Wikipedia, 2008)

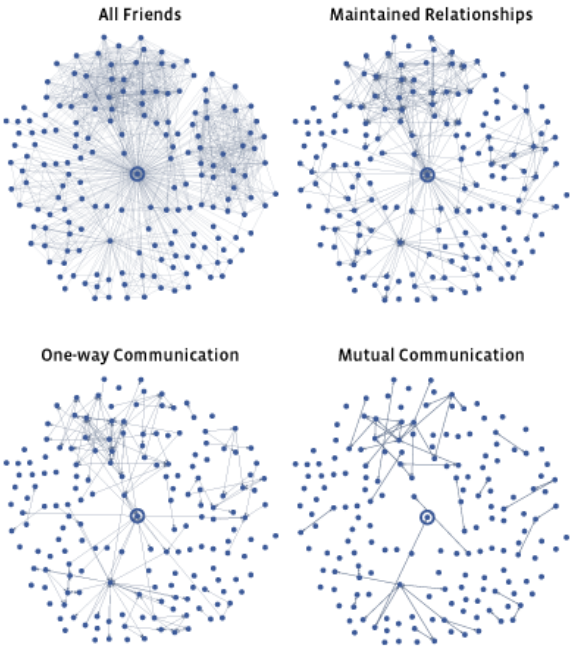

Figure 1: Visualization of a user's friends, where the extent of each type of relationship or communication is indicated by saturation (shade of blue) of the connection.

6 Evaluation

Although the components we use can be (and in most cases, have been) thoroughly evaluated in relative isolation, it is important to understand the consequences of their use in concert. Improvements to spam detection should be evident both in tests on annotated[6] data and in decreased reports or complaints from users.

User-generated metadata, in some cases a simple report of offensive content or a friend's compromised account, is a natural source of both labeled test data and training data. Our customer service processes are thus tightly integrated with machine learning efforts. See Figure 1 for communications in a small segment of the social graph.

7 Conclusion

Preliminary experiments with user-initiated machine translation of friend-generated content suggest it will soon be valuable. It is crucial to design this in a scalable way, such that it extends to arbitrarily many languages[7], both draws on and sup-

[6]Either a binary classification (spam or non-spam) or a gradient scale, possibly incorporating dimensions of phishiness, spamminess, or other types of solicitousness.

[7]Including underrepresented ones like Oshindonga.

ports our internationalization efforts, and should be useful on mobile devices (including in the spoken modality).

Our introductory questions (from Section 1) are far from fully answered, but we hope this work might help to address them.

1. The number and strength of connections, speed and frequency of communication, and diversity of languages individuals are exposed to all have strong influences on language change.

2. Stylistic variations in an individual's language are evident in that it can be more accurately captured as a mixture of models, each of which is suited to a specific situation, style, or set of interlocutors.

3. Two speakers is sufficient for a language. A small model can adequately describe a language, if each data point is a deviation from another language.

4. A common language is far from necessary for communication[8]. A set of arbitrary individuals' language models can be combined (and pruned, evolved) to derive the pidgin they might speak.

7.1 Future Work

Social natural language processing is (in a sense) in its infancy. We hope to capture aspects of its evolution, just as the field comes to better describe and understand ongoing changes in human languages. We have not yet satisfactorily answered our second question, but expect more fine-grained analyses to follow, using our framework to compare and contrast a variety of languages (from Bantu to Balinese) and phenomena (inside jokes, cross-linguistic usage of l33t and txt msg terms).

We hope to facilitate this by providing an API to allow researchers access to anonymized[9], aggregated data.

Acknowledgments

This technology is developed with support from i18n team (engineers, language managers and others) at Facebook, and all our international users. Thanks to our data scientists for the visualization of a user's friends, and the extent of communication connecting them.

[8] Photos, emoticons and tone of voice (for example) go a long way. We hope personalized (including speech-to-speech) translation will continue to bridge the language divide.

[9] Also honoring users' privacy settings.

References

Toine Bogers and Antal van den Bosch. 2008. Using language models for spam detection in social bookmarking. In *Proceedings of the ECML/PKDD Discovery Challenge*.

Dhruba Borthakur, 2007. *The Hadoop Distributed File System: Architecture and Design*. The Apache Software Foundation.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867.

David Ellis. 2009. A case study in community-driven translation of a fast-changing website. In *Proceedings of the 13th International Conference on Human-Computer Interaction HCII (to appear)*, San Diego, California, USA.

Dmitriy Genzel. 2005. *Creating Algorithms for Parsers and Taggers for Resource-Poor Languages Using a Related Resource-Rich Language*. Ph.D. thesis, Brown University.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Comput. Linguist.*, 30(4):417–449.

Ralf Schenkel, Tom Crecelius, Mouna Kacimi, Thomas Neumann, Josiane Parreira, Marc Spaniol, and Gerhard Weikum. 2008. Social wisdom for search and recommendation, June.

Er Strehl, Joydeep Ghosh, and Raymond Mooney. 2000. Impact of similarity measures on web-page clustering. In *In Workshop on Artificial Intelligence for Web Search (AAAI 2000*, pages 58–64. AAAI.

Wikipedia. 2008. Jaccard's similarity coefficient.

Shengliang Xu, Shenghua Bao, Yunbo Cao, and Yong Yu. 2007. Using social annotations to improve language model for information retrieval. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 1003–1006. CIKM.

Ding Zhou, Jiang Bian, Shuyi Zheng, Hongyuan Zha, and Lee C. Giles. 2008. Exploring social annotations for information retrieval. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 715–724, New York, NY, USA. ACM.