Fluency detection on communication networks

Tom Lippincott and **Benjamin Van Durme**
Human Language Technology Center of Excellence
Johns Hopkins University
tom@cs.jhu.edu, vandurme@cs.jhu.edu

Abstract

When considering a social media corpus, we often have access to structural information about how messages are flowing between people or organizations. This information is particularly useful when the linguistic evidence is sparse, incomplete, or of dubious quality. In this paper we construct a simple model to leverage the structure of Twitter data to help determine the set of languages each user is fluent in. Our results demonstrate that imposing several intuitive constraints leads to improvements in performance and stability. We release the first annotated data set for exploring this task, and discuss how our approach may be extended to other applications.

1 Introduction

Language identification (LID) is an important first step in many NLP pipelines since most downstream tasks need to employ language-specific resources. In many situations, LID is a trivial task that can be addressed e.g. by a simple Naive Bayes classifier trained on word and character n-gram data (Lui and Baldwin, 2012): a document of significant length will be quickly disambiguated based on its vocabulary (King et al., 2014). However, social media platforms like Twitter produce data sets in which individual documents are extremely short, and language use is idiosyncratic: LID performance on such data is dramatically lower than on traditional corpora (Bergsma et al., 2012; Carter et al., 2013). The widespread adoption of social media throughout the world amplifies the problem as less-studied languages lack the annotated resources needed to train the most effective NLP models (e.g. treebanks for statistical parsing, tagged corpora for part-of-speech tagging, etc). All of this motivates the research community's continued interest in LID (Zampieri et al., 2014).

Tweet #1	Коз эверисинг ю ду ис мэджик
Tweet #2	omg favourite day of the week!

Table 1: Multilingual social media users often communicate in different languages depending on their intended audience, such as with these Russian and English tweets posted by the same Twitter account

In this paper, we consider the closely-related task of determining an actor's *fluencies*, the set of languages they are capable of speaking and understanding. The observed language data will be the same as for LID, but is now considered to indicate a latent property of the actor. This information has a number of downstream uses, such as providing a strong prior on the language of the actor's future communications, constructing monolingual data sets, and recommending appropriate content for display or further processing.

This paper also focuses on the situation where a very small amount of content has been observed from the particular user. While this may seem strange considering the volume of data generated by social media, this is dominated by particularly active users: for example, 30% of Twitter users post only once per month (Leetaru et al., 2013). This content-starved situation is exacerbated by certain use-cases, such as responding to emergency events where sudden focus is directed at a particular location, or focusing on new users with shallow histories.

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1025–1029,
Austin, Texas, November 1-5, 2016.

2 Previous Work

Twitter and other social media platforms are a major area of ongoing NLP research, including dedicated workshops(NAA, 2015; ACL, 2014). Previous work has considered macroscopic properties of the entire Twitter network (Gabielkov et al., 2014), and pondered whether it is an "information" or "social" network (Myers et al., 2014). Studies have focused on determining user attributes such as gender (Li et al., 2015), political allegiance (Volkova et al., 2014), brand affinity (Pennacchiotti and Popescu, 2011a), sentiment analysis (West et al., 2014), and more abstract roles (Beller et al., 2014). Such demographic information is known to help downstream tasks (Hovy, 2015). Research involving social media communication networks has typically focused on *homophily*, the tendency of users to connect to others with similar properties (Barberá, 2014). A number of papers have employed features drawn from both the content and structure of network entities in pursuit of latent user attributes (Pennacchiotti and Popescu, 2011b; Campbell et al., 2014; Suwan et al., 2015).

3 Definitions

We refer to the entities that produce and consume communications as *Actors*, and the communications (packets of language data) as *Messages*. Each message occurs in a particular *Language*, and each actor has a set of *Fluencies*, representing the ability to produce and consume a message in a given language. We refer to a connected graph of such entities as a *Communication Network*. For Twitter data, messages are simply associated with a single actor, who is in turn associated with other actors via the "following" relationship, the actor's "friends" in Twitter's terminology.[1] We assume each message (tweet) is written in a single language, and actors are either fluent or not in each possible language.

[1]Note that, confusingly, Twitter's "friend" relationship is not symmetric: Mary's friends are users she has decided to follow, and not necessarily vise-versa.

4 Twitter Data Set

To build a suitable data set[2] for fluency detection, we first identified 1000 Twitter users who, according to the Twitter LID system, have tweeted in Russian and at least one additional language. For each of these "seed" users, we gather a local context (a "snowflake") as follows: we choose 20 of their friends at random. For each of these friends, we choose 15 of *their* friends (again, at random). Finally, we randomly pull 200 tweets for each identified user. The data set consists of 989 seed users, 165,042 friends, and 55,019,811 tweets. We preserve all Twitter meta-data for the users and tweets, such as location, follower count, hashtags, etc, though for the purposes of this paper we are only interested in the friendship structure and message text. We then had an annotator determine the set of languages each of the 1000 seed users is fluent in. For each seed user, the annotator was presented with their 200 tweets, grouped by Twitter language ID, and was asked to 1) flag users that appear to be bots and 2) list the languages they believe the user is fluent in. These steps are reflected in Figure 4. Over 50% (507) of the users were flagged as possible bots and not used in this study. The remaining 482 were observed employing 7 different languages: Russian, Ukrainian, German, Polish, Bulgarian, Latvian, and English. At most, a single user was found to be fluent in three languages.

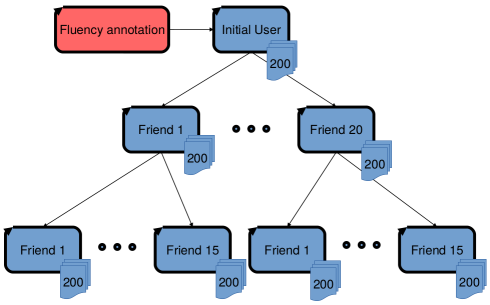

Figure 1: Structure of one snowflake in the Twitter Fluency data set.

5 Structure-Aware Fluency Model

Our goal was to explicitly model each actor's fluency in different languages, using a model with sim-

[2]The full data set is available at `www.university.edu/link`

ple, interpretable parameters that can be used to encode well-motivated assumptions about the data. In particular, we want to bias the model towards the belief that actors typically speak a small number of languages, and encode the belief that all actors participating in a message are highly likely to be fluent in its language. Our basic hypothesis is that, in addition to scores from traditional LID modules, such a model will benefit from considering the behavior of an actor's interlocutors. To test this, we designed a model that employs scores from an existing LID system, and compare performance with and without awareness of the communication network structure. To demonstrate the effectiveness of the model in situations with sparse or unreliable linguistic content, we perform experiments where the number of messages associated with each actor has been randomly down-sampled.

Linear Programming Linear Programming (LP) is a method for specifying constraints and cost functions in terms of linear relationships between variables, and then finding the optimal solution that respects the constraints. The restriction to linear equations ensures that the objective function is itself linear, and can be efficiently solved. If some or all variables are restricted to take discrete values, referred to as (Mixed) Integer Linear Programming (ILP), finding a solution becomes NP-hard, though common special cases remain efficiently solvable. We specify our model as an ILP with the hope that it provides sufficient expressiveness for the task, while remaining intuitive and tractable. Inference is performed using the Gurobi modeling toolkit (Gurobi Optimization, 2015).

Model definition Given a communication network with no LID information, ideally we would like to determine the language of each message, and the set of languages each actor is fluent in. Initially, we assume access to a probablistic LID system that maps unicode text to a distribution over possible languages. We use the following notation: $A_{1:T}$ and $M_{1:U}$ are the actors and messages, respectively. $F(a_i)$ is a binary vector indicating which languages we believe actor a_i is fluent in. $L(m_i)$ is a one-on binary vector indicating which language we believe message m_i is written in. $P(m_i)$ is the set of actors participating in message m_i: for Twitter data, where messages are (usually) not directed at specific users, we treat a user and the users' friends as participants. $LID(m_i)$ is a real vector representing the probability of message m_i being in each language, according to the LID system.

To build our ILP model, we iterate over actors and messages, defining constraints and the objective function as we go. There are two types of structural constraints: first, we restrict each message to have a single language assignment:

$$\sum L(m_i) = 1 \tag{1}$$

Second, we ensure that all actors participating in a given message are fluent in its language:

$$\forall a \in P(m_i), L(m_i) \times F(a) = 1 \tag{2}$$

The objective function also has two components: first, the *language fit* encourages the model to assign each message a language that has high probability according the the LID system:

$$\text{LF} = \sum_{m \in M} L(m) \times LID(m) \tag{3}$$

Second, the *structure fit* minimizes the cardinality of the actors' fluency sets (subject to the structural constraints), and thus avoids the trivial solution where each actor is fluent in all languages:

$$\text{SF} = -\sum_{a \in A} \sum F(a) \tag{4}$$

Finally, the two components of the objective function are combined with an empirically-determined *language weight* to get the complete objective function:

$$\text{LW} \times \text{LF} + (1.0 - \text{LW}) \times \text{SF} \tag{5}$$

Note that these are not all linear relationships: in particular, the multiplication operator cannot be used in ILP when the operands are both variables, as in equation 2. There are however techniques that can represent these situations in a linear program by introducing helper variables and constraints (Bisschop, 2015).

Language Identification Scores and Fluency Baseline To get LID scores, we ran the VaLID system (Bergsma et al., 2012) on each message, and normalize the output into distributions over 261 possible languages. VaLID is trained on Wikipedia data (i.e. out-of-domain relative to Twitter), although it does employ hand-specified rules for sanitizing tweet text, such as normalizing whitespace and removing URLs and user tags. VaLID uses a data-compression approach that is competitive with Twitter's in-house LID, despite no consideration of geographic or user priors. These language scores are used in the structure-aware model to compute the language fit.

Because VaLID makes no use of the communication network structure, we also use its scores to create a baseline structure-unaware fluency model. To get structure-unaware baseline scores for the fluency identification task, we average the LID distributions for each actor's messages and consider them fluent in a language if its probability is above an empirically-determined threshold.

Tuning parameters We empirically determine the thresholds for the baseline model and the language weights for the structure-aware model via a simple grid search, repeated 100 times. We randomly split the data into 20%/80% tune/test sets, and evaluate filter thresholds and language weights from 0 to 1 in .01 increments, with messages per actor ranging between 1 and 10. We expected the baseline model to have a consistent optimal threshold (though with higher performance variance with fewer messages), and this was borne out with optimal performance at a threshold of 0.06, independent of the number of messages per actor. For the structure-aware model, the optimal language weight was 0.9, although the entire range from 0.1–0.9 showed similar performance. This result was surprising, as we expect the structure-aware model to rely heavily on the structural fit when the number of messages is small, and on the language fit when the number is large. This trend doesn't emerge because the structural fit actually relies on the language fit to make assignments for the seed actor's friends and their messages.

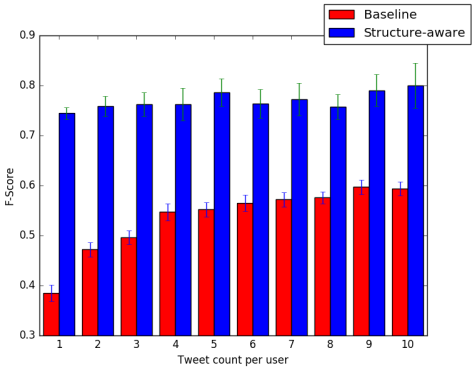

Figure 2: Performance of baseline and structure-aware models as a function of the number of messages per actor used as evidence. Each bar represents the average over 100 random tuning/testing splits, with whiskers showing the standard deviation.

6 Results and discussion

Figure 2 compares the performance[3] of the structure-aware ILP model with the baseline model as a function of the number of messages per actor, using the empirically-determined threshold and language weight. At the left extreme, the models only have a single, randomly-selected message from each actor. As this number increases, the baseline model improves as it becomes more likely to have seen enough messages to reflect the actor's full spectrum of language use. The structure-aware model is able to make immediate use of the actor's friends, immediately reaching high performance even when the language data is very sparse. Its most frequent type of error is over-hypothesizing fluency in both Ukrainian and Russian, when the user is in fact monolingual, followed by incorrectly hypothesizing fluency in English. This is understandable given the similarity of the languages in the former case, and the popularity of English expressions, titles, and the like in the latter.

7 Conclusion

We have presented promising results from leveraging structural information from a communication network to improve performance on fluency detection in situations where direct linguistic data is sparse. In addition to defining the task itself,

[3]F-score calculated based on correct and hypothesized fluency-assignments for each actor.

we release an annotated data set for training and evaluating future models. Planned future work includes a more flexible decoupling of the language and structure fits (in light of Section 5), and moving from pre-existing LID systems to joint models where LID scores are directly informed by structural information.

References

2014. ACL Joint Workshop on Social Dynamics and Personal Attributes in Social Media.

Pablo Barberá. 2014. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*, 23:76–91.

Charley Beller, Rebecca Knowles, Craig Harman, Shane Bergsma, Margaret Mitchell, and Benjamin Van Durme. 2014. I'm a belieber: Social roles via self-identification and conceptual attributes. In *Proceedings of the 52rd Annual Meeting of the Association for Computational Linguistics*, pages 181–186, Baltimore, Maryland, USA.

Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clayton Fink, and Theresa Wilson. 2012. Language identification for creating language-specific Twitter collections. In *Proc. Second Workshop on Language in Social Media*, pages 65–74.

Johannes Bisschop. 2015. Aimms optimization modeling.

W.M. Campbell, E. Baseman, and K. Greenfield. 2014. langid.py: An off-the-shelf language identification tool. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media*, pages 59–65, Dublin, Ireland.

Simon Carter, Wouter Weerkamp, and Manos Tsagkias. 2013. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. *Lang. Resour. Eval.*, 47(1):195–215, March.

Maksym Gabielkov, Ashwin Rao, and Arnaud Legout. 2014. Studying social networks at scale: Macroscopic anatomy of the twitter social graph. In *SIGMETRICS '14*, Austin, Texas, USA.

Inc. Gurobi Optimization. 2015. Gurobi optimizer reference manual.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 752–762, Beijing, China.

Ben King, Dragomir Radev, and Steven Abney. 2014. Experiments in sentence language identification with groups of similar languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 146–154, Dublin, Ireland.

Kalev Leetaru, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook. 2013. Mapping the global twitter heartbeat: The geography of twitter. *First Monday*, 18(5).

Shoushan Li, Jingjing Wang, Guodong Zhou, and Hanxiao Shi. 2015. Interactive gender inference with integer linear programming. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 2341–2347. AAAI Press.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 25–30, Jeju, Republic of Korea.

Seth A. Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. 2014. Information network or social network? the structure of the twitter follow graph. In *WWW '14 Companion*, Seoul, Korea.

2015. NAACL International Workshop on Natural Language Processing for Social Media.

Marco Pennacchiotti and Ana-Maria Popescu. 2011a. Democrats, republicans and starbucks afficionados: User classification in twitter. In *KDD '11*, San Diego, California, USA.

Marco Pennacchiotti and Ana-Maria Popescu. 2011b. A machine learning approach to twitter user classification. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 281–288.

Shakira Suwan, Dominic Lee, and Carey Priebe. 2015. Bayesian vertex nomination using content and context. *WIREs Comput Stat*, 7:400–416.

Svitlana Volkova, Glen Coppersmith, and Benjamin Van Durme. 2014. Inferring user political preferences from streaming communications. In *Proceedings of the 52rd Annual Meeting of the Association for Computational Linguistics*, pages 186–196, Baltimore, Maryland, USA.

Robert West, Hristo Paskov, Jure Leskovec, and Christopher Potts. 2014. Exploiting social network structure for person-to-person sentiment analysis. *Transactions of the Association for Computational Linguistics*, 2:297–310.

Marcos Zampieri, Liling Tang, Nikola Ljubešić, and Jörg Tiedemann. 2014. Discriminating similar languages shared task at coling 2014.